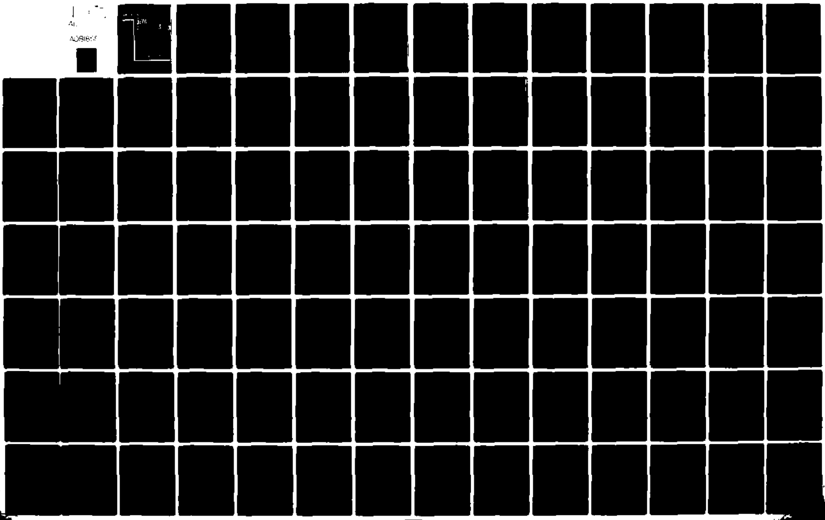
AD-A081 855 CONTROL DATA EDUCATION CO MINNEAPOLIS MN F/6 5/9
COMPUTERIZED INSTRUCTIONAL ADAPTIVE TESTING MODEL: FORMULATION --ETC(U)
FEB 80 S J KALISCH F33615-77-C-0071

UNCLASSIFIED AFHRL-TR-79-33 NL



AFHRL-TR-79-33

AIR FORCE

HUMAN RESOURCES LABORATORY

2

# COMPUTERIZED INSTRUCTIONAL ADAPTIVE TESTING MODEL: FORMULATION AND VALIDATION

By

Stanley J. Kalisch
Educational Testing Service
3445 Peachtree Road, N.E.
Atlanta, Georgia 30326

*UNDER SUBCONTRACT FROM*
Control Data Education Company
8100 34th Avenue, South
Minneapolis, Minnesota 55440

TECHNICAL TRAINING DIVISION
Lowry Air Force Base, Colorado 80230

February 1980

Final Report

Approved for public release; distribution unlimited.

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235

## NOTICE

When U.S. Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This final report was submitted by Control Data Education Company, 8100 34th Avenue, South, Minneapolis, Minnesota 55440, under contract F33615-77-C-0071, project 2313, with Technical Training Division, Air Force Human Resources Laboratory (AFSC), Lowry Air Force Base, Colorado 80230. Dr. Roger Pennell (TTY) was the Contract Monitor for the Laboratory.

This report has been reviewed by the Office of Public Affairs (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public, including foreign nations.

This technical report has been reviewed and is approved for publication.

MARTY R. ROCKWAY, Technical Director
Technical Training Division

RONALD W. TERRY, Colonel, USAF
Commander

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

## 19 REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

18 1. REPORT NUMBER
AFHRL-TR-79-33

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

6 4. TITLE (and Subtitle)
COMPUTERIZED INSTRUCTIONAL ADAPTIVE TESTING MODEL: FORMULATION AND VALIDATION

9 5. TYPE OF REPORT & PERIOD COVERED
Final rept.

6. PERFORMING ORG. REPORT NUMBER

10 7. AUTHOR(s)
Stanley J. Kalisch, Jr

15 8. CONTRACT OR GRANT NUMBER(s)
F33615-77-C-0071

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Control Data Education Company
8100 34th Avenue, South
Minneapolis, Minnesota 55440

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
61102F
16 2313T408
17 T4

11. CONTROLLING OFFICE NAME AND ADDRESS
HQ Air Force Human Resources Laboratory (AFSC)
Brooks Air Force Base, Texas 78235

11 12. REPORT DATE
February 1980

13. NUMBER OF PAGES
106

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)
Technical Training Division
Air Force Human Resources Laboratory
Lowry Air Force Base, Colorado 80230

12 107

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

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

18. SUPPLEMENTARY NOTES

This work was accomplished under subcontract by the Educational Testing Service, 3445 Peachtree Road, N.E., Atlanta, Georgia 30326.

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

adaptive testing
branched testing
computer simulation
instructional testing
mastery/non-mastery testing
objective mastery/non-mastery
tailored testing

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

The study included a formulation of eight versions of an adaptive testing model and computer simulations used to compare the accuracy and efficiency of the versions to each other and to conventional testing modes. The adaptive testing model was designed to be applicable to the needs and problems incurred in assessing mastery/non-mastery of instructional objectives by trainees. The overall purpose was to design a testing model that could provide greater accuracy in mastery/non-mastery classifications with fewer test items than are presented with conventional tests. The study included two phases of computer-simulated tests of the models. The first phase employed Monte Carlo simulations; the second used existing data obtained from trainees in a Weapons Mechanics training course. The results of the first phase showed that one of the versions was superior in accuracy to a

4/c 393 471

Item 20 Continued:

conventional testing procedure. Although the other versions provided the same accuracy as the conventional mode, selection of other error parameters would probably be able to increase the accuracy of the adaptive testing versions. All the adaptive versions required significantly fewer items to make mastery/non-mastery decisions. In the second phase, two of the adaptive versions were selected for comparison using data from actual trainees. The results showed the adaptive testing models could provide the same decisions as are made now by the Air Force in the Weapons Mechanics course, except that on the average 75 percent fewer test items would be needed. It was concluded that the adaptive testing model could be used to substantially reduce testing time and maintain the present level of accuracy in decision making, not only in Weapons Mechanics but also in other courses requiring the same types of decisions.

## SUMMARY

### Problem

In Air Force technical training, involving segments of systematic instruction, the frequency with which measurement occurs and the associated number of measurement items administered constitute large time demands with respect to the goal of minimizing training time. These time demands are often magnified in a self-paced training environment whenever the frequency of measuring students' performance is increased.

A self-paced training environment has been established in the Advanced Instructional System (AIS), developed by McDonnel Douglas Corporation under contract with the Air Force Human Resources Laboratory. AIS "is a comprehensive CMI/CAI [Computer-Managed Instruction/ Computer-Assisted Instruction] system for the administration and management of large scale individualized technical training at Lowry Air Force Base, Colorado" (McCombs, 1977, p.7). Trainee prescriptions are based in part on the results of tests measuring each trainee's performance on objectives of the course. A terminal is used to read a trainee's answer sheet, score the test forms, transmit data to a central computer, and print a prescription. The system is designed to handle a daily load of 2100 trainees for four Air Force technical training courses (Lamos & Waters, 1978).

In view of the fact that the Air Force has invested in a computerized training program with a high volume of testing, means of reducing the time for testing without a reduction in information obtained would be desirable. One of these means may be computerized adaptive testing. This form of testing provides a test to an individual based upon the person's prior responses to items. Only items for which adequate information can be obtained are presented. The process avoids presenting test items from which little or no additional information about the individual's performance can be obtained. Computerized adaptive testing can provide a method of significantly reducing the number of test items presented and therefore reduce testing time. Hence, adaptive testing has the potential of providing significant reduction in training time and dollars.

In addition to the potential of reducing testing time, an adaptive testing procedure should reflect the necessary accuracy required in making decisions about trainees' proficiencies. Adaptive testing procedures need to provide for different levels of decision-making accuracy as required for job performance or for subsequent phases of training.

One point that is generally overlooked, however, is that Air Force technical training is a nominally criterion-referenced system in which the acquisition of training objectives is evaluated using criterion-referenced knowledge and performance tests. Therefore, a testing system within this environment must be amenable to the requirements of, and problems inherent in, criterion-referenced testing.

## Approach

Eight versions of an adaptive testing model were formulated. The model was based upon the use of the Wald binomial probability ratio test and item prediction procedures employing a data base of test item responses collected from prior examinees.

The validation portion of the research effort was a two-phase study using computer simulations. The first phase employed Monte Carlo simulations. Data were generated to reflect test item response data that would be expected from actual trainees. The hypothetical examinees in the simulations were nested within one of the eight versions of the adaptive testing model or a control testing version. Each simulated examinee within a test mode went through 10 hypothetical tests using the assigned test mode.

The control testing version consisted of prespecified fixed numbers of items presented. For each objective, an examinee's mastery or nonmastery was based upon the individual's score in relation to a single cutting score.

The adaptive testing model requires the specification of parameters such as the error levels associated with false mastery and false non-mastery decisions. The prime investigator selected the levels that appeared to be applicable to such training situations as the Air Force Weapons Mechanics course. Other parameters specified were also selected to reflect the same training and job-related needs.

The primary purpose of the first phase was to compare the accuracy and the efficiency of the eight versions to each other and to the control version. Analysis of variance and a posteriori tests were used to make the comparisons. Accuracy was measured in terms of total loss--the sum of the losses incurred for incorrect decisions on each objective. A loss value is a positive or a zero number assigned to an action-outcome combination (Hays & Winkler, 1970). A zero loss value is assigned to any combination that reflects the best actions under the true circumstances. If an action is less desirable than the best actions, an error is associated with the action and is assigned a positive value reflecting the level of error involved. A total loss was computed for each examinee by test replication. The "true" mastery/non-mastery status was known for each examinee on each objective of each test since the data were generated by computer. Efficiency was measured in terms of the number of items presented.

The second phase employed two of the adaptive testing versions selected on the basis of their performance in the first phase. The primary purpose of the second phase was to compare the versions' efficiencies to each other and to the testing mode customarily used by the Air Force. This last mode consisted of the presentation of the same items for a given test to each examinee. Data used in the simulations were obtained from examinees in the Weapon Mechanics course on four different tests.

## Results

Phase I results showed that one of the adaptive testing versions provided greater accuracy than the control version. Although the other versions were equally as accurate as the control, the versions could be expected to have greater accuracy if the values for the error parameters were more stringently set. All versions of the adaptive testing model were more efficient than the control version

Phase II results replicated the Phase I results with regard to efficiency. Both adaptive testing versions made decisions more efficiently than the conventional testing mode. The adaptive versions required on the average 75 percent fewer items than the conventional testing version.

The overall results of the simulations suggest that a test of some versions of the model with real examinees within the actual training environment would not be premature. The results indicate the potential of adaptive testing for substantially reducing testing time.

The results also indicate additional appropriate research that may be conducted with computer simulations. Such areas as the effects of differential error levels on accuracy and efficiency should be pursued. These areas and others may be investigated relatively easily using the computer programs developed under this contract.

# PREFACE

A research effort such as the one being reported in this document required the efforts of many more people than the one listed as author. I wish to thank Dr. Roger Pennell of the Air Foıce Human Resources Laboratory and Art Neuman of Control Data Corporation for their suggestions toward the development and validation of the adaptive testing model. I also wish to express my sincerest thanks to Ruth MacInnes of Educational Testing Service for her perseverance in typing the document.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF ILLUSTRATIONS

Figure

# THE PURPOSE AND NATURE OF THE STUDY

Testing trainees within their training programs is a vital function in ascertaining who can perform sufficiently well or who needs additional training. As important as the function is that testing serves, it cannot require exorbitant amounts of time. Unfortunately, when there are numerous training objectives on which measures of performance are necessary, presenting even a few test items for each objective results in a test that is fairly lengthy. Compounding the problem is the realization that classifying performance on an objective as adequate or inadequate on the basis of a few items can be highly inaccurate.

One possible solution is to use adaptive testing, that is, tests with items that are selected for an examinee on the basis of that individual's prior responses or information on the individual. This type of testing avoids presenting items which provide little of no additional information about an examinee's performance. If an examinee has demonstrated inadequate performance on an objective with four items, there may be no need to present more items for this objective. If an examinee has shown proficiency on an objective that incorporated skills bearing upon subordinate objectives, there may be no need to present items on these subordinate objectives. In essence, adaptive testing provides an individualized test that can adequately ascertain performance on objectives in an efficient manner.

The purpose of this study was to formulate and validate an adaptive testing model addressing accuracy and efficiency of computerized instructional testing in Air Force technical training. The model and the validation studies were designed in relation to the type of testing used in typical training programs such as the Weapons Mechanics Training Course (63ABR46230) conducted at Lowry Air Force Base, Colorado. Although the model was designed to be relevant to Air Force needs, the model is applicable to many other computer-based testing situations.

Two types of computer simulations were employed. A Monte Carlo simulation was used to compare the accuracy and efficiency of eight versions of the adaptive testing model and a control treatment consisting of tests with randomly selected items. The two versions that appeared to be somewhat better than the others were then used in a second computer simulation with data previously collected in the Weapons Mechanics course. The purpose of the second simulation was to judge how efficient the versions of the model were when actual trainee data were used. The relative efficiency of the versions was compared to the efficiency level attained by the present method of testing at Lowry AFB.

## THE ADAPTIVE TESTING MODEL

The adaptive testing model developed combines the models of Ferguson (1969) and Kalisch (1974a, 1974b). Ferguson's procedure employs the Wald binomial probability ratio test to determine mastery/nonmastery of hierarchically interrelated objectives. Kalisch's procedure employs a process that predicts item responses based upon prior examinees' data. In this study, a combination of obtained and predicted item responses was used with the Wald binomial probability ratio test and hierarchical configurations of objectives to ascertain each examinee's mastery/non-mastery of objectives.

### Ferguson's Adaptive Testing Model

Ferguson (1969) applied an adaptive model that assessed an individual's proficiency on an objective by using the Wald binomial probability ratio test after each response by the examinee. (A detailed description of the Wald probability ratio test, Ferguson's application of the procedure, and issues related to the application of the Wald procedure to criterion-referenced testing appear in Appendix A.) Once a mastery/non-mastery decision regarding the individual's proficiency on the objective had been reached, a branch to another objective was based on the mastery/non-mastery decision and the proportion of errors made. If the examinee mastered an objective, branching was directed to a superordinate objective. The step size from the objective tested to a superordinate objective increased as the number of correct responses on the tested objective increased. If the examinee did not master an objective, branching was directed to a subordinate one. The step size from the objective tested to a subordinate one increased as the number of correct responses to the tested objective decreased.

Ferguson used a validated hierarchy of skills. "A sufficient but not necessary element for branched testing is a valid hierarchy of skills upon which routing may be based. The lack of a valid hierarchy does not affect construction or administration of the branched test; however, it has a profound effect on the usefulness of the results derived from the test" (Ferguson, 1969, p.87).

### Kalisch's Adaptive Testing Models

Kalisch (1974b) used adaptive models that predicted the responses to unanswered items of a test by comparing the item response patterns of examinees to the response patterns of subjects whose item response data had been previously obtained. The item having a probability closest to 0.5 of being answered correctly was presented to the simulated examinee. The procedure continued until all the item responses had been obtained or predicted.

Although Kalisch used the responses to items corresponding to instructional objectives, no determination of objective mastery/non-mastery was made. Kalisch suggested that the same fundamental procedure be used and mastery/non-mastery decisions be based upon the proportion of items to which correct responses were given or to which predictions indicate that a correct response is probable.

## The Adaptive Testing Model in This Study

Like Kalisch's previous procedure, a vector of obtained and predicted responses was used. The vector of correct and incorrect responses was applied to the Wald binomial probability ratio test to ascertain mastery/non-mastery of objectives. As with Ferguson's procedure, the specification of the hierarchical configuration of the objectives indicated for which objectives mastery may be assumed and for which ones further testing is necessary to ascertain mastery or non-mastery.

## A General Description of the Adaptive Testing Model Formulated in this Study

Figure 1 shows a hypothetical configuration of objectives. Objective 5 has Objectives 2 and 3 as its immediate subordinates, or prerequisites. This means that mastery of the skill or competency represented by Objective 5 requires that both Objectives 2 and 3 be mastered. Non-mastery of either or both Objectives 2 and 3 implies non-mastery of Objective 5. The figure indicates no prerequisite to Objective 2. Objective 1 is prerequisite to both Objectives 3 and 4. The immediate prerequisites to Objective 6 are Objectives 2, 3, and 4. No prerequisites are indicated for Objective 7.

Generally some objectives are considered more important or critical. Other objectives may be subordinate or prerequisite to the former objectives, those of primary concern. If mastery can be ascertained for the "objective of primary concern," then there appears to be little, if any, need to assess performance on the subordinate objectives. If one wanted to use the model to assess performance on all the objectives, then every objective would be identified as an objective of primary concern.

The model first requires the specification of a hierarchical configuration of the objectives and the identification of the "objectives of primary concern." Terminal objectives would generally be a part of this classification (e.g., Objectives 5, 6, and 7 in Figure 3). Prerequisite objectives may also be considered terminal objectives. For example, if Objective 1 is a prerequisite to numerous objectives within a course, it may be that the objective is of sufficient importance to be of primary concern. As another example, understanding of technical terminology in a subject area may not only

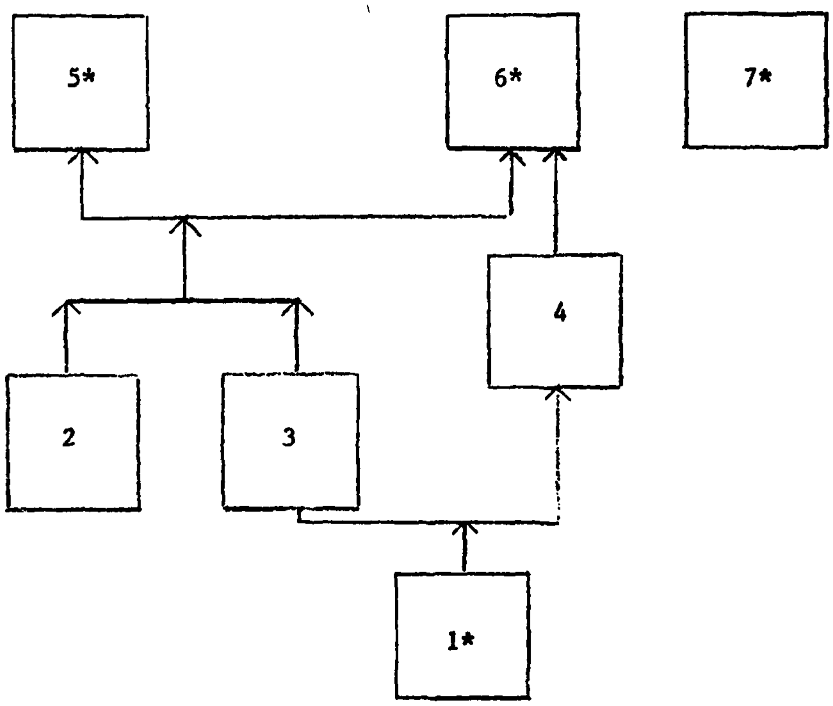


Figure 1. Hypothetical Hierarchical Configuration of Objectives

NOTE: * indicates an "objective of primary concern."

be prerequisite to other skills, but also considered important as an end result. Hence, this objective may also be of primary co.icern.

The mastery and non-mastery levels of each objective are then specified or determined. Methods for setting these values are described in the section "Setting $\theta_1$ and $\theta_2$" in Appendix A.

The model developed for this study has two methods of item selection. Both methods consider item mastery and non-mastery, that is, whether an item has been answered correctly or incorrectly, respectively. With the first method, an item having the highest mastery/non-mastery agreement with the objectives of primary concern is selected for presentation to an examinee. With the second method, an item having the highest mastery/non-mastery agreement with each item corresponding to the objectives of primary concern. The measures of agreement for each method are based on data obtained from prior examinees.

Regardless of the item selection procedure used, responses obtained for the selected items are matched with the same patterns displayed by prior examinees. The item response matching technique provides information both for selection of the next item and prediction of responses to unpresented items.

On the basis of both types of responses, obtained and predicted, mastery/non-mastery classifications of objectives are made using the Wald probability ratio test. For example, suppose that for an objective, responses were obtained to three items and predictions were made for six other item responses. These nine responses (correct/incorrect for each item) are then to be used in the following formula:

$$S = (R \cdot \log_{10} \frac{C_f}{C_p}) + [(N - R) \cdot (\log_{10} \frac{1 - C_f}{1 - C_p})]$$

where R = number of items answered (or predicted as being answered) correctly

N = number of items (number presented plus the number predicted)

$C_f$ = the critical non-mastery score (difficulty of the objective for non-masters)

$C_p$ = the critical mastery score (difficulty of the objective for masters)

In Appendix A it is shown that

$$S = \log_{10} \frac{P_1(d,n)}{P_0(d,n)}$$

where $n$ = the number of elements in the sample
$d$ = the number of defective elements in the sample.

$P_0$ and $P_1$ are determined from the formulas

$$P_0\ (d,n) = \theta_0^{\,d}\,(1 - \theta_0)^{n-d}$$

$$P_1\ (d,n) = \theta_1^{\,d}\,(1 - \theta_1)^{n-d}\ ,$$

where $\theta_0$ = maximum proportion of defectives permitted for the entire collection

$\theta_1$ = proportion of defectives at or above which the collection is to be rejected.

Changing the context from "defectives" to "mastery," we note that

$$\begin{aligned} n &= N \\ d &= N - R \\ \theta_0 &= 1 - C_p \\ \theta_1 &= 1 - C_f \end{aligned}$$

Hence,

$$\begin{aligned} P_0\ (d,n) &= P_0\ (N - R,\ N) \\ &= \theta_0^{N - R}\ (1 - \theta_0)^{N-(N - R)} \\ &= \theta_0^{N - R}\ (1 - \theta_0)^{R} \\ &= (1 - C_p)^{N - R} \cdot C_p^{\,R} \end{aligned}$$

and similarly,

$$\begin{aligned} P_1\ (d,n) &= P_1\ (N - R,\ N) \\ &= (1 - C_f)^{N - R} \cdot C_f^{\,R} \end{aligned}$$

Therefore,

$$S = \log_{10} \frac{C_f{}^R \ (1 - C_f)^{N - R}}{C_p{}^R \ (1 - C_p)^{N - R}}$$

$$= \log_{10} \left[ \frac{C_f}{C_p}^R \cdot \frac{1 - C_f}{1 - C_p}^{N - R} \right]$$

$$= R \log_{10} \frac{C_f}{C_p} + (N - R) \log_{10} \frac{1 - C_f}{1 - C_p}$$

As an example, suppose two items were answered correctly, one was answered incorrectly, and the predictions indicated a correct response to each of the six other items. Hence, $N = 9$ and $R = 8$. Suppose $C_f$ and $C_p$ have been determined from prior data. Let $C_f = .6$ and $C_p = .9$. Based on these values

$$S = (8 \log_{10} \frac{.6}{.9}) + [(9-8) (\log_{10} \frac{.4}{.1})]$$

$$S \approx -1.409 + .602$$

$$S \approx -0.807$$

Mastery/non-mastery classifications are determined by comparing the value of S to ratios involving $\alpha$ and $\beta$ (classification error parameters) as follows:

- If $S \geq \log_{10} \frac{1 - \beta}{\alpha}$, the objective is not mastered

- If $S \leq \log_{10} \frac{\beta}{1 - \alpha}$, the objective is mastered

- If neither of the above conditions is true, no mastery/non-mastery classification is possible (additional item responses are necessary).

Suppose that the values selected for $\alpha$ and $\beta$ are .1 and .2, respectively. Then

$$A = \log_{10} \frac{1 - \beta}{\alpha} \approx .903$$

$$B = \log_{10} \frac{\beta}{1 - \alpha} \approx -.653$$

Since $S < \beta$, the objective is classified as mastered.

The model developed for this study has two response matching procedures. One method matches the pattern of dichotomously scored (correct/incorrect) responses for an examinee with response patterns of prior examinees. The second method employs an additional matching procedure on the basis of mastery/non-mastery classifications of objectives. An example serves to illustrate the two methods. Suppose that an examinee has answered 10 items, six correctly and four incorrectly. Additionally, suppose that it has been concluded that the examinee has not mastered one of the objectives of primary concern. With the first method, prior subjects' data with exactly the same pattern of correct and incorrect responses to the same items presented to the examinee are used to make predictions on other item responses and to select additional items for presentation to the examinee. With the second method, prior subjects' data must not only have the same pattern of correct and incorrect responses but also have the same mastery and non-mastery classifications on the same objectives as the examinee.

In addition to the two item selection methods and the two response matching procedures, the model developed has another dichotomous option--whether obtained responses should or should not be checked for examinee inconsistency. Inconsistent responses by an examinee would include (a) correctly answering an item by guessing when other related items have been answered incorrectly and (b) incorrectly answering an item by mistake. The inconsistency check may require the presentation of more items but would be expected to improve the accuracy of item response predictions and item selection and hence improve the accuracy of mastery/non-mastery classifications.

The model assumes that mastery of an objective implies mastery of all its immediate subordinate objectives; non-mastery of an objective implies neither mastery nor non-mastery of the immediate subordinates. Mastery classification on an objective of primary concern results in an assumption that all the immediately prerequisite or subordinate objectives are mastered, unless a subordinate is also of primary concern. Non-mastery classification on an objective of primary concern results in testing each immediate subordinate as an objective of primary concern.

The model assumes that the classification of an objective for which insufficient items exist for a mastery/non-mastery decision is "indeterminate." This decision occurs whenever the pool of available items is exhausted before a mastery/non-mastery decision can be made. Such an objective is presently treated as "unmastered" although this could be altered without affecting other components of the model. Rather than assuming the objective to be unmastered, the process could ascertain which classification zone the examinee's response pattern more closely approached. Ferguson (1969) used this procedure, but only after asking for 30 item responses for the objective. If a trainee cannot demonstrate mastery performance within a realistically expected number of items, immediately prescribing remedial instruction appears to be more efficient than giving a lengthy test to make a decision. An objective for which an undesirably high proportion of "indeterminate" classifications has been made indicates an insufficient number of items, insufficient item discrimination, or unrealistically high specifications for acceptable misclassification errors.

The adaptive testing procedures defined are terminated when either of the following conditions occurs:

1. All objectives have been classified as mastered or unmastered.

2. The number of prior examinee observations in the data base upon which predictions are based is less than two.

For the first condition, the test is terminated. For the second condition, unpresented and unpredicted items corresponding to objectives of concern are randomly presented to the examinee. Termination of the test occurs when each objective is classified. (Some objectives will be classified as "indeterminate" if the item pool for the objective is exhausted without a mastery/non-mastery decision possible.)

The overview of the procedures in the adaptive testing model formulated for this study has described eight cells of a 2 x 2 x 2 configuration of options. These derive from three options, each with two conditions:

1. Two methods of item selection based upon

   A. Item-objective agreement
   B. Inter-item agreement

2. Two response matchino procedures based upon

   A. Only item response patterns
   B. Both item response and objective classification patterns

3. A dichotomous option regarding examinee response inconsistency.

The details, including mathematical formulas, for these options are described in succeeding paragraphs, but since other elements of the model affect the computations, a discussion of these elements follows.

Basing Decisions on the Data Base. The decisions made in the adaptive testing process are dependent upon information collected from prior examinees. Although the model presently assumes that each prior examinee has answered all the items for each objective, it could accommodate a data base consisting of responses by prior examinees to overlapping subsets of item pools. Decisions such as selection of items for presentation and prediction of correctness/incorrectness of item responses are made on the basis of the interrelation of item responses by prior examinees whose response patterns match the present examinee's pattern. For each item response obtained from an examinee using the adaptive test, a smaller subset of prior subjects' data is used to make decisions--a subset of examinees' dichotomously scored responses, exactly like the present examinee's response pattern.

As an example of the ways item response patterns are used, consider the dichotomously scored item responses in Table 1. The "1's" indicate correct responses, whereas "0's" represent incorrect responses. These data would be obtained from prior examinees.

Table 1

Data Base of Item Responses

| Subjects | Items | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| 1 | 1 | 1 | 0 |
| 2 | 1 | 1 | 1 |
| 3 | 1 | 1 | 0 |
| 4 | 1 | 1 | 1 |
| 5 | 1 | 1 | 0 |
| 6 | 1 | 0 | 1 |
| 7 | 0 | 1 | 0 |
| 8 | 1 | 0 | 1 |
| 9 | 0 | 0 | 0 |
| 10 | 1 | 0 | 0 |

Suppose that an examinee is presented with Item 2. The item is answered correctly. Based upon the data base in Table 1, the conditional probability of answering item 1 is computed as follows:

$$P\ (C_1|C_2) = \frac{P(C_1 \cap C_2)}{P(C_2)} = \frac{\frac{5}{10}}{\frac{6}{10}} = \frac{5}{6}$$

where $P(C_1|C_2)$ means the probability of answering Item 1 correctly given that Item 2 has been answered correctly

$C_1$ means Item 1 is answered correctly

$C_2$ means Item 2 is answered correctly.

Similarly, to determine the probability that the examinee will answer Item 3 correctly given that Item 2 has been answered correctly, the same procedure is used. Hence

$$P\ (C_3|C_2) = \frac{P(C_3 \cap C_2)}{P(C_2)} = \frac{\frac{2}{10}}{\frac{6}{10}} = \frac{1}{3}$$

If the probability of answering an item is sufficently high or low, one should be willing to predict that the examinee will answer the item correctly or incorrectly, respectively. The test used in the adaptive model is that if the probability is less than or equal to $\alpha$, where $\alpha$ is the Type I error level for the corresponding objective, item non-mastery is predicted. Similarly if the probability is greater than or equal to $1-\beta$, where $\beta$ is the Type II error for the objective, item mastery is predicted.

For example, if $\alpha = .2$ and $\beta = .1$,

$$P\ (C_1|C_2) = \frac{5}{6} \leq 1 - \beta = .9$$

and $$P\ (C_3|C_2) = \frac{1}{3} \leq 1 - \beta = .9.$$

Hence no predictions of item mastery could be made. Similarly,

$$P\ (C_1|C_2) = \frac{5}{6} \geq \alpha = .2$$

and $P\ (C_3|C_2) = \frac{1}{3} \geq \alpha = .2.$

Therefore no predictions of item non-mastery would be made.

If item 3 were presented and the examinee answered the item correctly, the probability of answering item $C_1$ correctly would be computed as follows:

$$P(C_1|C_2,\ C_3) = \frac{P(C_1 \cap C_2 \cap C_3)}{P(C_2\ C_3)} = \frac{\frac{2}{10}}{\frac{2}{10}} = 1.$$

Since $P(C_1|C_2,\ C_3) = 1 > 1 - \beta = .9$, it would be predicted that Item 1 would be answered correctly by the examinee.

Response pattern matching is also used for selecting items for presentation. Item selection procedures are described in an ensuing section.

The minimum number of examinees from whom item responses must be obtained for the data base is not known. Kalisch (1974b) showed that, in the versions of the model he used in an earlier study, 100 to 200 examinees' sets of responses to 33 items, produced results that were not significantly inferior to those obtained when large data bases (2000 sets of responses) were used. Kalisch did not make mastery/non-mastery classifications for the objectives. His criteria for selection of items were based on the difficulties of items rather than the interrelationship of an item to the objectives or to other items. Because his versions are different from those developed for this study, his conclusions do not indicate a required data base size applicable to the present model.

Individual Examinee's Historical Data Record

Although decisions may be made solely on the basis of data collected from prior examinees, such decisions do not consider systematic patterns demonstrated by the examinee using adaptive testing. In situations such as training programs in which testing occurs at numerous points in the program, there is an opportunity to obtain information on each individual. This information can be used to determine how well the adaptive process is working and whether it should be altered for the individual; the information can also be used to determine whether the prediction criteria should be changed for the individual, to increase accuracy or efficiency in decision making. To capitalize on the systematic patterns demonstrated by individuals, a data record for each examinee is updated after each test. When sufficient data are contained in the record, these data are used for item prediction and selection purposes. This collection of data is to be called an examinee's historical data record.

The historical data record for an individual permits transforming the conditional difficulty of an item, based upon the data base, into the individual's difficulty level. For example, on the basis of an examinee's item response pattern and information in the data base, the conditional difficulty of an item might equal .75. This means that for prior examinees with the same item response pattern, 75 percent answered the item correctly. On the basis of this information, the probability that the present examinee would answer the item correctly equals .75. Using the individual's historical data might indicate a different probability. Suppose that over a period of time data have been collected on the individual's actual performance in correctly/incorrectly answering items with group conditional difficulties approximately equal to .75. For example, the historical data for an individual may indicate the person's probability of answering an item with a group conditional difficulty value of .75 is .8. On the basis of these individual difficulty estimates, the adaptive testing procedure is modified to increase accuracy and/or efficiency.

Lacking a generalizable continuous function that would adequately define the transformation of group conditional difficulties into individual difficulties, the collection of data within difficulty intervals was used. Each individual's historical data record is subdivided into 21 intervals of group conditional difficulties. The intervals are defined in Table 2. For each of the 21 categories, the following data are collected:

1. Present mean value (difficulty)

2. Number of observations upon which mean is based

3. Lowest value observed in the interval.

Table 2

Definitions of Group Conditional Difficulty Intervals

| Interval Number | Values in Interval |
|---|---|
| 1 | [ 0, .025) |
| 2 | [ .025, .075) |
| 3 | [ .075, .125) |
| 4 | [ .125, .175) |
| 5 | [ .175, .225) |
| 6 | [ .225, .275) |
| 7 | [ .275, .325) |
| 8 | [ .325, .375) |
| 9 | [ .375, .425) |
| 10 | [ .425, .475) |
| 11 | [ .475, .525) |
| 12 | [ .525, .575) |
| 13 | [ .575, .625) |
| 14 | [ .625, .675) |
| i5 | [ .675, .725) |
| 16 | [ .725, .775) |
| 17 | [ .775, .825) |
| 18 | [ .825, .875) |
| 19 | [ .875, .925) |
| 20 | [ .925, .975) |
| 21 | [ .975, 1.000] |

Note: In this table and throughout the text a bracket indicates the number next to it is included in the interval, whereas a parenthesis indicates the number next to it is not included in the interval. Hence the interval [.025, .075) contains all values between .025 and .075, including .025, but not including .075.

Suppose the following conditional difficulties of items were observed as well as whether the examinee answered the item correctly or incorrectly, (1 = correct; 0 = wrong):

| Observation | Conditional Difficulty | Right or Wrong |
|---|---|---|
| 1 | .74 | 0 |
| 2 | .75 | 1 |
| 3 | .76 | 1 |
| 4 | .728 | 0 |
| 5 | .771 | 1 |

All of the conditional difficulties fall in the interval [.725, .775)--Interval 16. Hence, for these five observations, Interval 16 would have the following data stored:

- Mean = .6 (3 out of 5 items answered correctly--sum 0 and 1's, and divide by 5)
- Number of observations = 5
- Lowest conditional difficulty observed = .728

If a sixth observation applicable to Interval 16 is incurred, the values are updated. Hence, suppose an item is answered incorrectly when its conditional difficulty is .726, the revised data would be as follows:

- Mean = .5
- Number of Observations = 6
- Lowest Observed value = .726

Data are collected for an individual's historical data record on every item presented and on randomly selected items for which predictions have been made. For each item selected for presentation there exists a group conditional difficulty determined from the subset of data base responses having exactly the same item response pattern as the present examinee. The item response is dichotomously scored and the results recorded in the historical data record. For an item on which a prediction has been made, there is also a group conditional difficulty value. Based upon the number of observations already collected within the corresponding conditional difficulty interval, a decision is made whether a response is to be collected on the predicted item. If a response is obtained, the data are added to the historical record but are not used for decision making during the present adaptive test.

The collection of responses to items on which predictions have been made may decrease the efficiency without increasing the accuracy of the decisions made during the examinee's present adaptive test. As sufficient data are compiled in the individual's historical record, the need to present such items would decrease. The model assumes that historical data collection never ceases, but diminishes as the data in the historical record increases. Hence, the reduction in efficiency with no increase in accuracy during the present adaptive test is permitted in order to obtain data that will increase accuracy and/or efficiency on future adaptive tests.

In this study an algorithm for deciding whether an item with a predicted response should also be presented, was defined.

The probability $P_n$ that an item with a predicted response should also be presented was defined as

$$P_n = \begin{cases} \dfrac{N}{N + 3} & \text{if } N < 10 \\ \dfrac{3}{N + 1} & \text{if } N \geq 10 \end{cases}$$

where N = number of observations presently in the corresponding group conditional difficulty interval.

A random number r in the closed interval [0, 1] is selected. If $r \leq P_n$, the item is also presented to the examinee. The significance of 10 observations in the formula is based upon a subjective decision that 10 observations would be minimally required to estimate the individual's item difficulty value within any single group conditional difficulty interval.

One way in which the historical data are used is to obtain difficulties more accurately representing the individual's difficulty level than can be obtained directly from group conditional difficulties. The computation of the individual's item difficulty for item i is represented by F(i = 1) whereas the group conditional difficulty for the same item is represented by P(i = 1). Suppose that P(i = 1) = .87. In Table 2 Interval 18 contains the probability for answering items with difficulties in the interval [.825, .875). If the mean value for the interval is based upon 10 or more observations, the mean is assigned to F(i = 1). If Interval 18 were to have fewer than 10 observations but each of the two adjacent categories (in this case Intervals 17 and 19) has 10 or more observations, then an estimate of the mean for Interval 18 would be computed as follows:

$$\frac{[(n_{17} + n_{19}) \times (\frac{\text{mean}_{17} + \text{mean}_{19}}{2})] + [n_{18} \times \text{mean}]}{n_{17} + n_{18} + n_{19}}$$

where $n_{17}$, $n_{18}$, $n_{19}$ represent the number of observations for Intervals 17, 18, 19, respectively. If an interval has neither 10 or more observations nor two adjacent categories with 10 or more observations each, then the interval has insufficient historical data. In this case F(i = 1) is set equal to P(i = 1).

Another way in which the historical data are used is to alter the difficulty value for which predictions of item response correctness and incorrectness are made. The item responses obtained from an examinee are matched against sets of responses obtained from prior examinees. From data base response vectors exactly matching the examinee's response vector, predictions of other item responses (i.e., whether the item will be answered correctly or incorrectly) are made. The decisions to predict item response correctness and incorrectness are dependent on the values of $\alpha$ (false non-mastery) and $\beta$ (false mastery) for the objective to which the item corresponds. The complements serve as the critical values for item predictions. If $F(i = 1) \geq 1 - \alpha$, it is predicted that Item i will be answered correctly. If the probability of answering an item incorrectly $F(i = 0) > 1 - \beta$, it is predicted that Item i will be answered incorrectly. Note that $F(i = 0)$ equals $1 - F(i = 1)$.

Item Selection

Two methods of item selection were developed for the model. The first method considers the mastery/non-mastery agreement between each item not presented or predicted with all the objectives of primary concern. The second method considers the mastery/non-mastery agreement between each item not presented or predicted with all the unpresented and unpredicted items corresponding to objectives of primary concern. (Item mastery or non-mastery means that the item has been answered correctly or incorrectly, respectively.)

Item-Objective Agreement. A coefficient of agreement is computed for each item not presented and for which prediction of correctness/incorrectness has not yet occurred. The item with the highest coefficient is presented to the examinee. The formula for the coefficient of agreement between Item i and the n objectives of primary concern is

$$C_{(i;O_1,O_2,\ldots O_n \mid \vec{r},\vec{s})} =$$

$$[ \{ \sum_{u=1}^{n} [\text{Prob } (O_u = 1) \mid (\vec{r},\vec{s},i = 1)] \bullet [F(i = 1) \mid \vec{r},\vec{s}] \}$$

$$+ \{ \sum_{u=1}^{n} [\text{Prob } (O_u = 0) \mid (\vec{r},\vec{s},i = 0)] \bullet [F(i = 0) \mid \vec{r},\vec{s}] \} ] /n$$

where i is the item under consideration
$O_1$, $O_2$, $\ldots O_n$ are the n objectives of concern
$i = 1$ means Item i is answered correctly
$i = 0$ means Item i is answered incorrectly
$O_u = 1$ means Objective u is mastered
$O_u = 0$ means Objective u is not mastered
$\vec{r}$ is the vector of objective mastery/non-mastery classifications for the examinee
$\vec{s}$ is the vector of examinee's dichotomously scored item responses
$F(i = 1)$ is the probability of the examinee's answering the item correctly
$F(i = 0)$ is the probability of the examinee's answering the item incorrectly.

The formula is devoid of the objective vector $\vec{r}$ if the response matching procedure is based on only item response patterns.

Inter-Item Agreement. The coefficient of agreement of an item with other items indicates the extent to which the correctness/incorrectness of the item response agrees with the correctness/incorrectness of responses to the other items. A coefficient of inter-item agreement is computed for each item neither presented nor predicted. The item with the highest coefficient is presented to the examinee. The formula for the coefficient of agreement between item $i_x$ and the n other items corresponding to objectives of concern is

$$A(i_x;i_1,i_2,\ldots i_n) = [\ \{ \sum_{j=1}^{n} [P(i_j = 1)|(\vec{r},\vec{s},i_x = 1)]\cdot[F(i_x = 1)|\vec{r},\vec{s}]\ \}$$
$$+ \{ \sum_{j=1}^{n} [P(i_j = 0)|(\vec{r},\vec{s},i_x = 0)]\cdot[F(i_x = 0)|\vec{r},\vec{s}]\ \} /n$$

where $P(i_j = 1)$ is the group conditional difficulty of item $i_j$ (the probability of answering the item correctly)
$P(i_j = 0)$ is the probability of answering item $i_j$ incorrectly $[P(i_j = 0) = 1 - P\ (i_j = 1)]$
$F(i_x = 1)$ is the individual's probability of answering item $i_x$ correctly
$F(i_x = 0)$ is the individual's probability of answering item $i_x$ incorrectly.
$\vec{r}$ is the objective mastery/non-mastery pattern for the examinee.
$\vec{s}$ is the item response pattern (correct/incorrect) for the examinee.

As in the item-objective method, if the response matching procedure is based solely on the item responses, then the objective vector $\vec{r}$ is not included in the formula.

## Response Matching Procedures

Two response matching procedures were defined. With the first method, a vector $\vec{s}$ of dichotomously scored responses is generated for an examinee. With each additional response collected within an examination, the $\vec{s}$ vector increases. The individual's $\vec{s}$ vector is matched with sets of responses in the data base. Only data base sets with exactly the same $\vec{s}$ vector (the same pattern of ones and zeros to exactly the same questions answered by the examinee) are considered.

With the second method, not only is the $\vec{s}$ vector used but also an $\vec{r}$ vector of mastery/non-mastery classifications for objectives is employed. Only data base sets with exactly the same $\vec{s}$ and $\vec{r}$ vectors are considered. With both methods the matching procedure provides the subset of data base entries that is used for making predictions and selecting other items for presentation.

Examinee Response Inconsistencies

"Untrue" responses by an examinee are those responses that do not agree with the examinee's "true" response--that is, the examinee's response that is not arrived at by guessing and has not been erroneously selected or created. "Untrue" responses are expected to occur in such cases as:

- Selecting the correct answer by guessing when in actuality overall performance indicates the examinee should have answered the item correctly.

- Providing an incorrect answer because of misinterpretation of a phrase in the question.

- Pressing answer choice "2" on a terminal keyboard when the examinee intended to press "3", when "2" is the correct choice.

Item responses which are provided by an examinee but which are contrary to the examinee's "true" response introduce potential measurement error into any testing process. In the adaptive test model, erroneous responses introduce error into $\vec{s}$, the item response vector. Predictions of other item responses and selection of items for presentation are based on $\vec{s}$. Generally it would be expected that item prediction errors would affect the accuracy of the system, whereas errors in item selection would reduce the efficiency of the system. Prediction and selection errors would occur since the adaptive testing process relies on matching the examinee's $\vec{s}$ with exactly the same response vectors in the data base. Errors introduced into $\vec{s}$ would produce a comparison between the examinee's performance and the wrong subset of prior examinees. Even if some of the response sets in the data base contain the same errors as made by the present examinee, it would be expected that for each item the majority of prior examinees have provided responses that concur with their "true" responses. Hence errors introduced into the examinee's item response vector would be expected to compare the examinee's performance to an inappropriate subset of prior examinees.

The adaptive testing model has included an optional component that checks for potentially "untrue" responses by comparing the examinee's inter-item response consistency to the inter-item response consistency demonstrated by all prior examinees whose data are included in the data base. When this option is selected, it is necessary that at least two items be presented for the examinee's responses prior to making predictions or other item selections based on the item response vector $\vec{s}$. The present model requires that a set of items be independently selected and presented. The number of items must be sufficient so that the probability of answering all of them correctly by chance alone is less than or equal to .05. This criterion was selected by the investigator based upon his judgment that it is reasonable. For multiple choice items each with five alternatives, the requirement is two items. For items each with three alternatives the requirement is three items. For items with two alternatives the requirement increases to five items.

The purpose of obtaining responses to a set of independently selected items is to determine whether the examinee has demonstrated a sufficiently consistent response pattern to warrant this pattern serving as the item response vector. A coefficient of relative interrelationship $R_x$ between item x and all other items for which responses have been obtained is computed as follows:

$$R_x = \frac{\sum_i G(x,i)}{\sum_i I(x,i)}, \quad \text{where}$$

$$G(x,i) = \begin{cases} 1 & \text{if both responses to Item x and Item i are correct or if both responses are incorrect.} \\ 0 & \text{if one response is correct and the other is wrong.} \end{cases}$$

$$I(x,i) = \{ [ \Sigma P(i = 1 \mid x = 1)] \cdot P(x = 1) \}$$

$$+ \{ \Sigma P(i = 0 \mid x = 0)] \cdot P(x = 0) \} .$$

G(x,i) is computed on the basis of the examinee's responses to Item x and all the other items presented. Suppose that three items have been scored as shown in Table 3.

Table 3
Three Scored Item Responses

| Item | Scored Response[1] |
|---|---|
| 1 | 1 |
| 2 | 1 |
| 3 | 0 |

[1] 1 = Right
0 = Wrong.

For item i = 1,

G(1,2) = 1 since both items were answered correctly
G(1,3) = 0 since one was answered correctly and the other was answered incorrectly.

I(i,x) is determined from the data base. The formula, although expressed differently from the G(x,i) is computed in the same manner. I(i,x) is the frequency of agreements in correctness or incorrectness of Items x and i, divided by the total number of cases. Suppose, for example, the data base consisted of the scored responses shown in Table 4.

Table 4

Scored Responses in Data Base

| Item | Examinees | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 1 | 1 | 1 | 1 | 0 | 0 |
| 2 | 1 | 0 | 0 | 0 | 0 |
| 3 | 1 | 1 | 0 | 0 | 0 |

I(1,2) = 3/5 = .6 (3 agreements out of 5 cases)
I(1,3) = 4/5 = .8 (4 agreements out of 5 cases)

On the basis of the above computations,

$$R_1 = \frac{\Sigma\, G(1,i)}{\Sigma\, I(1,i)} = \frac{G(1,2) + G(1,3)}{I(1,2) + I(1,3)}$$

$$= \frac{1 + 0}{.6 + .8}$$

$$= \frac{1}{1.4} \approx .714$$

In similar fashion

$$R_2 = \frac{\Sigma\, G(2,i)}{\Sigma\, I(2,i)} = \frac{G(2,1) + G(2,3)}{I(2,1) + I(2,3)}$$

$$= \frac{1 + 0}{.6 + .8} \approx .714$$

$$R_3 = \frac{\Sigma\, G(3,i)}{\Sigma\, I(3,i)} = \frac{G(3,1) + G(3,2)}{I(3,1) + I(3,2)}$$

$$= \frac{0 + 0}{.8 + .8} = 0$$

$R_x$ indicates the examinee's consistency as compared to that of prior examinees. It is possible that the present examinee demonstrates greater consistency than prior examinees, but when the present examinee's consistency is less than that for prior examinees, it is assumed that the person's item response pattern contains "untrue" responses. The criterion for sufficiently consistent responses by an examinee was set in this study at .90. If this criterion is not attained for each item, the item with the lowest $R_x$ value is temporarily removed from consideration as a member of the item response vector $\vec{s}$. Prior to making decisions based on $\vec{s}$, the item response vector must contain at least the required minimum number of elements (equal to the number of items to be answered to insure that the probability of guessing the correct answers is less than or equal to .05). If $\vec{s}$ contains fewer elements, other items must be independently selected without reference to $\vec{s}$. Whenever the number of elements in $\vec{s}$ equals or exceeds the minimum requirement, item selections and predictions are based upon $\vec{s}$. After the presentation of each additional item, all items for which responses are obtained are included in the calculations of the $R_x$ values. Hence, although an item response may be questioned and not included in $\vec{s}$, a future recalculation may indicate the item response to be consistent with the examinee's other responses. Likewise, items once contained in $\vec{s}$ may on a future recalculation be excluded.

## Eight Versions of the Adaptive Testing Model

Table 5 provides a delineation of the options used for each of the eight versions of the adaptive testing model that result from the three options, namely:

- Method of item selection
- Method of response matching
- Option of checking response inconsistencies

Table 5

Options Employed in the Eight Versions of the Adaptive Testing Model

| Version | Item Selection | Response Matching[1] | Inconsistency Check |
|---|---|---|---|
| 1 | Item-Objective | only $\vec{s}$ | No |
| 2 | Inter-Item | only $\vec{s}$ | No |
| 3 | Item-Objective | both $\vec{r}$ and $\vec{s}$ | No |
| 4 | Inter-Item | both $\vec{r}$ and $\vec{s}$ | No |
| 5 | Item-Objective | only $\vec{s}$ | Yes |
| 6 | Inter-Item | only $\vec{s}$ | Yes |
| 7 | Item-Objective | both $\vec{r}$ and $\vec{s}$ | Yes |
| 8 | Inter-Item | both $\vec{r}$ and $\vec{s}$ | Yes |

[1] $\vec{s}$ is the item response vector
$\vec{r}$ is the objective mastery/non-mastery classification vector

## PHASE I DATA GENERATION AND EXPERIMENTAL DESIGN

The purpose of Phase I of the study was twofold:

1. To test for the relative accuracy and efficiency of the eight versions of the adaptive testing model.

2. To study elements of the model to determine how well they are working, namely:

   - The relation of loss tc individuals' achievement levels

   - The nature and effectiveness of the individuals' historical data bases.

Accuracy was examined in terms of correct mastery/non-mastery classifications. Efficiency was investigated in terms of the number of items presented to examinees.

The control treatment to which the adaptive testing versions were being compared involved the testing of every objective. For each objective, a prespecified number of items was randomly selected for each examinee. Under the control version, examinees generally receive different items for an objective, but each receives the same number of items. For each objective, a randomly selected integer between three and six, inclusive, was chosen for the number of items to be presented. Mastery of an objective was obtained if an examinee obtained a score of N - 1 or higher, where N equals the number of items presented. A score of less than N - 1 resulted in a non-mastery classification. The resulting lengths of the tests and the mastery criteria reflected the parameters used in the Air Force Weapons Mechanics training program at Lowry AFB.

Phase I employed Monte Carlo simulations. Item response data were generated for hypothetical examinees who were to demonstrate some consistency in performance across examinations. This assumes that individuals in instructional programs demonstrate a certain consistent performance in mastering or not mastering objectives. The adaptive testing model uses each examinee's past test performance data to make decisions on present and future testing. Hence the Phase I experiment was designed to study the effects on efficiency and accuracy as past examinee test performance data became available.

For each examination by adaptive test version, two sets of examinee data were generated--one representing past examinees' responses and the other representing responses that are obtainable from present examinees. For the control version, only one set of examinee data was generated for each examination. A set of examinee responses was generated in two steps using two computer programs, GENTAB and GENRESP. For each examinee, GENTAB produced values for elements of consistency to be demonstrated across examinations. These elements were the examinee's achievement level and risk of guessing. The values from GENTAB and additional parameters were used to produce item responses through program GENRESP. Parameters specified for GENRESP included the following:

- Hierarchical configuration of the objectives.
- Objective parameters such as difficulty, discrimination, and passing criteria.
- Proportion and type of hierarchical errors.
- Guessing factor for answering items correctly.

Program GENTAB

A four-digit number representing the probability of mastering an objective was randomly selected in the open interval (0,1). The examinee records were sorted on the basis of achievement levels. Achievement levels ranged from zero through one, inclusive.

The second parameter generated for each examinee was the risk value for guessing. Based upon the assumption that individuals differ as to the risk they will take in guessing at the answers to items, a random number in an interval was assigned to the individual. The risk value represents the probability of attempting to guess the correct answer given that the examinee's true response to the item is incorrect. The value does not represent the probability of answering the item correctly, but rather the probability that a guess will be attempted. The random number chosen for an individual was in a prespecified interval. For this study, the following ranges were used with the proportion of examinees being assigned values in each of these categories (levels of guessing):

| Category | Probability of Guessing | Proportion of Examinees |
|---|---|---|
| 1 | 0.95 - 1.00 | .90 |
| 2 | 0.85 - 0.949 | .05 |
| 3 | .75 - .849 | .05 |

The high probability values reflect the non-penalty for guessing in the Air Force instructional testing environment.

Program GENRESP

For each set of data produced by GENTAB, parameters were specified to produce responses that included errors attributed to examinees' guessing and hierarchical inconsistencies among objectives. Program GENRESP was used to produce item responses for both the data bases (the hypothetical data obtained from prior examinees) and the test bases (the hypothetical data for the examinees taking the adap tive tests).

An Example of Item Response Generation. From program GENTAB, a file containing records, one for each hypothetical examinee, was generated. Each record contained an examinee's identification number, achievement level, and risk of guessing. An example of such a file appears in Table 6.

Table 6

Sample File Produced by GENRESP

| ID | Probability of Passing | Risk of Guessing |
|---|---|---|
| 001 | .98 | .96 |
| 002 | .96 | .87 |
| 003 | .92 | .99 |
| . | | |
| . | | |
| . | | |
| 049 | .09 | .78 |
| 050 | .05 | .95 |

The records were ordered in decreasing order of the achievement level. The same GENRESP file was used to produce item responses for all the examinees in any one testing version. Hence a file such as in Table 6 was used to generate examinees' responses to 10 different examinations, each of which was taken by the hypothetical examinees listed in the GENRESP file.

To generate the examinees' responses to an examination, the number of objectives and the number of items for each objective were specified. For all tests considered in this phase there were five objectives--two superordinate and three subordinate objectives. Each superordinate objective had 15 items and each subordinate had 20.

Hierarchical Configuration. The configuration was randomly determined. A subordinate objective could be subordinate either to one or both of the two superordinates. One configuration that was used is shown in Figure 2. Objective 3 is subordinate to both Objectives 1 and 2; Objective 4 is subordinate only to Objective 1; Objective 5 is subordinate only to Objective 2.

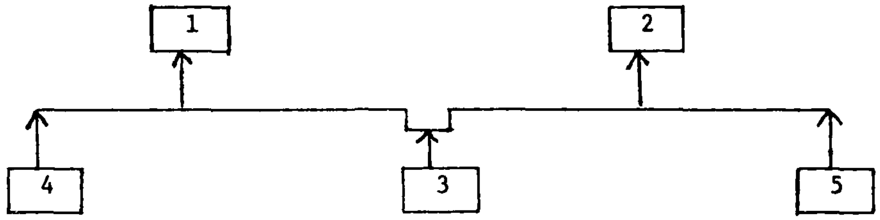


Figure 2. Randomly configured hierarchy of objectives.

Objective Difficulties. For each superordinate objective a random number in the interval [.75, .95] was selected. The objective difficulty is the average difficulty of the objective's items. The interval [.75, .95] reflects the average difficulty of items in the Weapons Mechanics course. Similarly for subordinates, a random number was selected in the same interval except that the value had to be greater than or equal to all its superordinate's difficulties. The rationale for this requirement is that superordinates are generally more difficult tasks than any one of the subordinates, hence the subordinate's difficulty would not be expected to be less than its superordinate's. Table 7 shows the objective difficulties selected for the example.

Table 7

Examples of Parameters Required for Program GENRESP

| | | | | Difficulties | |
|---|---|---|---|---|---|
| Objective | Objective Difficulty | Proportion Passing | Objective Discrimination | Masters | Non-Masters |
| 1 | .92 | .85 | .30 | .96 | .66 |
| 2 | .81 | .81 | .28 | .86 | .58 |
| 3 | .93 | .99 | .27 | .93 | .66 |
| 4 | .95 | .90 | .12 | .96 | .84 |

Proportion Passing. For each superordinate objective, a random number in the interval [.75, .90] was selected. The proportion passing the objective is synonymous with proportion mastering the objective. For subordinate objectives, a random number in the interval [.85, .99] was selected, provided the value equaled or exceeded all its superordinate's proportion passing values. Table 7 shows the proportion passing value for the example.

Objective Discriminations. Objective discrimination is define as the difficulty for masters minus the difficulty for non-masters Maximal discrimination would be obtained if all true masters answe all the items correctly and all true non-masters answered all the items incorrectly. No data for determining discriminations were available on the Weapons Mechanics examinations. Hence the ranges selected for the discrimination values were conjectures. For each superordinate, a random number in the interval [.2, .5] was selected. For each subordinate, the value selected was in the interval [.1, .4]. Table 7 shows the discrimination values selected for the example.

Difficulties for Masters and for Non-Masters. The difficulties of the objectives for non-masters were calculated from the formula

$$\text{difficulty}_{NM} = \begin{matrix}\text{objective}\\ \text{difficulty}\end{matrix} - \begin{matrix}\text{proportion}\\ \text{passing}\end{matrix} \times \begin{matrix}\text{objective}\\ \text{discrimination}\end{matrix}$$

The difficulties for masters was tnen calculated as follows:

$$\text{difficulty}_{M} = \text{difficulty}_{NM} + \begin{matrix}\text{objective}\\ \text{discrimination}\end{matrix}$$

For Objective 1 the calculations would be performed as follows:

$$\begin{aligned}\text{difficulty}_{NM} &= .92 - (.85 \times .30)\\ &= .665\\ \text{difficulty}_{M} &= .665 + .30\\ &= .965\end{aligned}$$

Table 7 shows the difficulties (rounded to the nearest hundredth) for masters and for non-masters.

Assignment of Mastery/Non-Mastery Status to Examinees. Proceeding sequentially through the file produced by GENTAB, the computer selected a random number in the interval (0,1) was made for each examinee. If the random number equaled or exceeded the examinee's achievement level, a mastery classification on the objective was given. This procedure was used independently for each of the two superordinate objectives. The process continued until the proportion of examinees mastering the objectives equaled their respective proportion passing values. Each subordinate objective to a mastered superordinate was classified as mastered also. This assumes a valid hierarchy, that is, mastery of superordinate objectives implies mastery of their subordinates. For the subordinate objectives, if the proportion passing was not yet attained, the file was sequentially processed for each examinee who had not yet been given mastery status. The random selection of a number in the interval (0,1) was compared to the examinee's achievement level. The subordinate objectives were classified using the same process as for the superordinates.

Hierarchical Errors. Thus far the process has assumed a completely valid hierarchy--one in which mastery of a superordinate implies mastery of all its subordinates for every examinee. Program GENRESP provides for hierarchical error levels to be specified. For the Phase I study, an error level between .00 and .10, inclusive, was randomly selected for each superordinate-subordinate relation. For the example under consideration the following error levels were selected:

| | | |
|---|---|---|
| Objectives 1 and 3 | — | 4% |
| Objectives 1 and 4 | — | 3% |
| Objectives 2 and 3 | — | 5% |
| Objectives 2 and 5 | — | 2% |

In the case of Objectives 1 and 4, three percent of the examinees who mastered Objective 1 did not master subordinate Objective 3. For Objective 3, the resulting hierarchical error rates would be expected to be higher than the four percent. Since the hierarchical errors for Objectives 1 and 3 are handled independently of Objectives 2 and 3, the resulting error between Objectives 1 and 3 would be expected to be greater than four percent but less than nine percent (the sum of the error rates involving the common subordinate).

Standard Deviations for Mastery/Non-Mastery Objective Scores. The item response generation algorithms used in GENRESP are based on the item difficulty and the examinee's objective score. The objective scores were selected randomly for an examinee based on a truncated normal distribution for the individual's mastery or non-mastery group, whichever was applicable. Two parameters used to generate either distribution are the objective difficulty for the group and the standard deviation of the scores. A random number in the interval (0, $\sqrt{d(1-d)}$) was selected, where d equals the applicable objective difficulty, either for masters or non-masters. Note that $\sqrt{d(1-d)}$) is the largest possible standard deviation for a distribution with a mean of d.

Assignment of Objective Scores to Examinees. For each objective, two truncated normal distributions were formed--one for masters and the other for non-masters. The distributions were defined by the applicable objective difficulties and standard deviations. Each distribution was truncated at a value midway between the difficulties for masters and non-masters. In the example, for each master a random number was selected in the interval [0.81, 1.00] so that frequency of scores selected would follow the truncated normal distribution for the mastery group. Similarly for each non-master, a value was selected in the interval [0.00, 0.81] to follow the defined non-mastery frequency distribution.

Generation of Item Difficulties. For each objective the mean item difficulties for masters or non-masters equals the objective difficulty for the masters and non-masters, respectively. The following assumptions were used in generating the item difficutlies for each objective:

- Item difficulties for masters are rectangularly distributed with a mean equal to the objective difficulty for masters and the maximum difficulty equal to 1.00.

- The item discrimination is equal to the objective discrimination value. (Therefore all items for an objective are equally discriminating.)

Hence for Objective 1 the item difficulties for masters ranged from 0.92 through 1.00; for non-masters the range was from 0.62 through 0.70.

Generation of Examinees' True Item Responses. For each objective each item response for an examinee was based on a probability of answering the item correctly. The algorithm used was

$$P(u = 1) = \begin{cases} d + \dfrac{\theta - \bar{\theta}}{1 - \bar{\theta}}(1 - d) & \text{if } \theta \geq \bar{\theta} \\[2ex] d + \dfrac{\theta - \bar{\theta}}{\bar{\theta}}\, d & \text{if } \theta < \bar{\theta} \end{cases}$$

where $P(u = 1)$ = the probability of answering the item correctly.

$\theta$ = examinee's objective score

$\bar{\theta}$ = mean objective score of the corresponding mastery/non-mastery group.

A random number r in the closed interval [0,1] was selected. If $r \leq P(u = 1)$, the examinee was assigned a correct item response; otherwise an incorrect item response was assigned.

Inclusion of Examinee Error. The factor of successful guessing was included in GENRESP. The probability that an attempt would be made to guess the correct answer, given that the examiner's "true" response would be incorrect, was derived by the formula

$$P_1 = g1(1 - \theta d)$$

where $g_1$ is the risk factor for the examinee (from GENTAB)

$\theta$ is the examinee's objective score

d is the item difficulty for the examinee's mastery or non-mastery group.

A random number $r_1$, in the interval [0, 1] was selected. If $r_1 \leq P_1$, the examinee would attempt to guess the correct answer. The probability of guessing correctly was obtained from the formula

$$P_2 = g2 + g_2\theta d$$

where $g_2$ is the guessing factor for the item (the probability of randomly selecting the correct answer)

$\theta$ and d are the same as defined previously.

For all items $g_2$ was set equal to .2, assuming five alternatives to each item. A random number $r_2$ in the interval [0,1] was selected. If $r_2 \leq P_2$, the examinee was credited with answering the item correctly.

Questions to be Answered

The intent of Phase I was to answer the following questions:

1. How do each of the eight versions of the adaptive testing model compare with regard to accuracy and efficiency?

2. Which version or versions of the model are superior in accuracy and efficiency to the others?

Although both accuracy and efficiency were to be addressed, the primary emphasis was on accuracy. Efficiency was a secondary issue in this phase of the study.

Since the accuracy of mastery/non-mastery classifications can be expected to increase as the α and β levels are reduced, it would have been possible to set the error levels so that greater accuracy could assuredly be obtained by the adaptive testing procedures than by the control. But low levels of α and β would have required the presentation of more items than required by the control treatment. This would have been undesirable since the overall goal of the study was to formulate an adaptive testing model that could provide the same or greater accuracy with fewer items presented. Hence, the α and β levels were selected to reflect what would be reasonable for the Air Force Weapons Mechanics course and the adaptive testing results were compared with respect to both the number of correct decisions and the number of items presented. Each of the following results would demonstrate the superior accuracy of the adaptive testing model:

- Adaptive tests provided more correct decisions and required fewer items than the control.
- Adaptive tests provided more correct decisions and required equally as many items to be presented as the control.

The following result would imply that the adaptive model could be the more accurate treatment, but would not definitively show its superiority:

- Adaptive and control treatments made equally as many correct decisions but the adaptive tests required fewer items.

The following results would demonstrate the superior accuracy of the control version:

- Control version provided more correct decisions and required fewer items than the adaptive test.
- Control version provided more correct decisions and required equally as many items to be presented as the adaptive tests.

The following results would imply that the control version could be more accurate, but would not definitively show its superiority:

- Control version and adaptive versions made equally as many correct decisions but the control required fewer items.

The following result would have demonstrated the two testing versions as equally accurate:

- Both versions made equally as many correct decisions with equally as many items presented.

The following results would have produced no conclusion regarding superior accuracy:

- Adaptive versions maue fewer correct decisions but required fewer items than the control.

- Control version made fewer correct decisions but required fewer items than the adaptive versions.

The most desirable result would be the adaptive test versions' demonstration of superior decision making with fewer items presented than the control version. This would demonstrate the adaptive testing model to be superior in accuracy and efficiency. Since there is no direct way to determine the existence of the α and β values that will show the adaptive testing versions to be superior in accuracy and efficiency, the values were arbitrarily selected. This approach could have resulted in not being able to conclude whether the adaptive model or the control version was superior in accuracy.

α and β Values Selected

In courses of medium or high criticality, it would be expected that the β level would be more stringently set than the α level. For skills involving safety and cost, the results of falsely classifying individuals as masters can be highly undesirable. A relatively high α level might result in unnecessarily providing remedial training (falsely classifying a master as a non-master), but this would not be as serious as the former error. Hence, the β level would be expected to be lower than α.

In some situations it would be appropriate for the α level to be lower than the β level. For example, programs in which the graduated trainee is to perform with skills acquired on the job under the close supervision of a superior rather than under the trainee's own cognizance may not require stringently set β levels. Errors due to false mastery classifications in the training program may very easily be corrected on the job. Also, if minimization of the training time is of high importance, then it would be desirable to minimize the number of false non-mastery decisions. Hence, for some programs α levels would be more stringently set than the β levels.

It should be recalled that as the levels of α and β are lowered, the number of items to be presented in order to make mastery/non-mastery decisions increases. Overly stringent criteria for α and β defeat testing efficiency.

In the Phase I study the values of α and β were set at .2 and .1, respectively. These values were selected by the investigator.

Experimental Design

Separate split-plot factorial analyses of variance were conducted for each of three dependent variables. The two independent variables were test version (eight versions of the adaptive testing model and the control version) and examination (10 replications). The three dependent variables were:

- Total loss associated with errors in mastery/non-mastery classifications.
- Total number of items presented.
- Number of items presented for item prediction purposes.

Total loss. A loss value is a positive or zero number assigned to an action-outcome combination (Hays & Winkler, 1970). A zero loss value is assigned to any combination that reflects the best actions under the true circumstances. If an action is less desirable than the best actions, an error is associated with the action and is assigned a positive value reflecting the level of error involved. The loss values appearing in Tables 8 and 9 represent the relative amounts of loss attributed to each mastery/non-mastery/indeterminate decision made given the "true" mastery/non-mastery status. The loss values were supplied by subject matter experts at Lowry AFB. The loss values presented in Tables 8 and 9 were developed using the instructions in Appendix C.

In Table 8 one sees that under the known true situation of mastery, the best decision is to classify performance on an objective as mastery. The positive numbers for decisions of "non-mastery" and "indeterminable" indicate there are errors involved with these decisions--the greater error being associated with the latter.

Table 8

Matrix of Loss Values Provided for
Objectives of Primary Concern

| Classification Decision | True Classification | |
|---|---|---|
| | Mastery | Non-Mastery |
| Mastery | 0 | 10 |
| Non-Mastery | 5 | 0 |
| Indeterminable | 7 | 3 |

Table 9

Matrix of Loss Values Provided for
Objectives of Secondary Concern

| Classification Decision | True Classification | |
|---|---|---|
| | Mastery | Non-Mastery |
| Mastery | 0 | 6 |
| Non-Mastery | 4 | 0 |
| Indeterminable | 5 | 2 |

Total loss equals the sum of the separate losses incurred for each objective decision for an examinee.

Total Number of Items Presented. Items for the adaptive tests were presented for the following reasons:

1. To provide information for predicting correctness/incorrectness of other items.

2. To obtain information solely for the examinee's historical data record.

The sum of these numbers equals the total number of items presented.

Number of Items Presented for Item Prediction Purposes. The first area listed above represents the third dependent variable. It was analyzed becuase the number of additional items presented solely for the historical data may be changed without altering the adaptive testing model and not affecting the number of items necessary for item prediction purposes.

Experimental Model. The split-plot factorial model used was

$$X_{ijkm} = \mu + A_i + B_j + \pi_{k(i)} + AB_{ij} + B\pi_{jk(i)} + \epsilon_{m(ijk)}$$

where $X_{ijkm}$ is the dependent variable

$A_i$ is the testing version effect

$B_j$ is the examination effect

$\pi_{k(i)}$ is the subject effect

A Posteriori Tests. With regard to the treatment effect, the Dunnett's *t* statistic was computed for each adaptive testing version with the control treatment. This a posteriori test was used for each dependent variable regardless of the *F* value obtained using the analysis of variance (Winer, 1971, p. 201). Therefore, each version was compared with the control version.

For other effects, Newman-Keuls tests were performed only when significant *F* values ($\alpha = .05$) were obtained from the analyses of variance.

Sample Size. Each data base from which predictions were made was composed of 300 *Ss'* sets of responses. For each of the 90 testing version by cells, 50 hypothetical examinees were used.

## PHASE I RESULTS

The results of the first phase of the study showed that all of the adaptive testing versions were significantly more efficient than the control. Only one version demonstrated significantly smaller losses than the control version. This was the sixth version listed in Table 5--adaptive testing using inter-item agreement, the item response vector, and the inconsistency check. Losses were greater for examinees with lower general achievement levels than those in the middle or higher levels. Among the adaptive versions, none required significantly fewer items. The adaptive testing versions varied in their rank orderings across test replications with regard to the number of items presented.

Elements of the model were studied from the Phase I results. The study indicated that the historical data record element was not working sufficiently well to be useful within the 10 replications. The results implied that the size of the data bases might be reduced but needed to include more response sets representing the more poorly performing examinees.

### Total Loss.

An analysis of variance indicated significant examination and testing version by examination effects ($\alpha = .05$). A quasi-*F* statistic was computed for the testing versions since the mixed effects model did not directly provide a mean sums of squares estimate for the required denominator (Winer, 1971 pp. 375-378). Table 10 shows the results of the analysis of variance. Tables 11 and 12 provide the descriptive statistics for the main effects.

The use of Hartley's test for homogeneity of variance (Winer, 1971 pp. 207-208) resulted in a rejection of the equal variance assumption. Hence, a more conservative test proposed by Box (Winer, 1971, p. 206) was used. The degrees of freedom corresponding to each numerator was reduced to one. The examination effect remained significant at the .05 level, but the testing version by examination interaction did not. Since the statistical test is extremely conservative, a graph of the interaction is presented in Figure 3.

Dunnet's test indicated that the only testing version significantly different ($\alpha = .05$) from the control was the sixth testing version--adaptive testing using inter-item agreement, the item response vector, and the inconsistency check. Although the seventh testing version's obtained *t* value did not exceed the critical value, the difference in the two was extremely small. The losses obtained for both testing versions were extremely close. The seventh testing version was the adaptive version using item-objective agreement based on both item response and objective classification vectors and employing the inconsistency check.

Table 10

Analysis of Variance For Total Loss

| Source | SS | df | MS | F |
|---|---|---|---|---|
| Between Subjects | 236577.96 | 449 | | |
| Testing Version | 4986.59 | 8 | 623.32 | 1.14 |
| Subjects within Groups | 231591.37 | 441 | 525.15 | |
| Estimates for quasi-_F_ calculations | | 457 | 544.624 | |
| Within Subjects | 91508.90 | 4050 | | |
| Examination | 6792.61 | 9 | 754.74 | 36.61* |
| Testing Version X Examination | 2886.56 | 72 | 40.09 | 1.94** |
| Examination X Subjects within Groups | 81829.73 | 3969 | 20.62 | |

*p < .01

**p < .01 for df(72,3969); p < .25 for df(1,3969).

Table 11

Descriptive Statistics of Total Loss For
Each Examinee per Testing Version

| | Total Loss | | |
|---|---|---|---|
| Testing Version | Mean | S. D. | Range |
| 1 | 5.84 | 10.25 | [0, 52] |
| 2 | 5.52 | 9.56 | [0, 48] |
| 3 | 5.88 | 9.79 | [0, 52] |
| 4 | 5.39 | 9.25 | [0, 60] |
| 5 | 5.03 | 8.46 | [0, 46] |
| 6 | 4.73 | 8.63 | [0, 49] |
| 7 | 4.84 | 8.55 | [0, 50] |
| 8 | 5.05 | 8.40 | [0, 44] |
| Control | 8.40 | 8.45 | [0, 60] |

Table 12

Descriptive Statistics of Total Loss For Each Examinee per Examination

| Examination | Total Loss | | |
|---|---|---|---|
| | Mean | S. D. | Range |
| 1 | 5.34 | 8.99 | [0, 52] |
| 2 | 7.45 | 11.17 | [0, 49] |
| 3 | 3.96 | 7.62 | [0, 38] |
| 4 | 4.48 | 6.90 | [0, 35] |
| 5 | 5.60 | 9.41 | [0, 46] |
| 6 | 6.45 | 10.33 | [0, 52] |
| 7 | 5.94 | 9.38 | [0, 44] |
| 8 | 7.80 | 9.17 | [0, 60] |
| 9 | 4.86 | 8.33 | [0, 50] |
| 10 | 4.40 | 8.37 | [0, 52] |

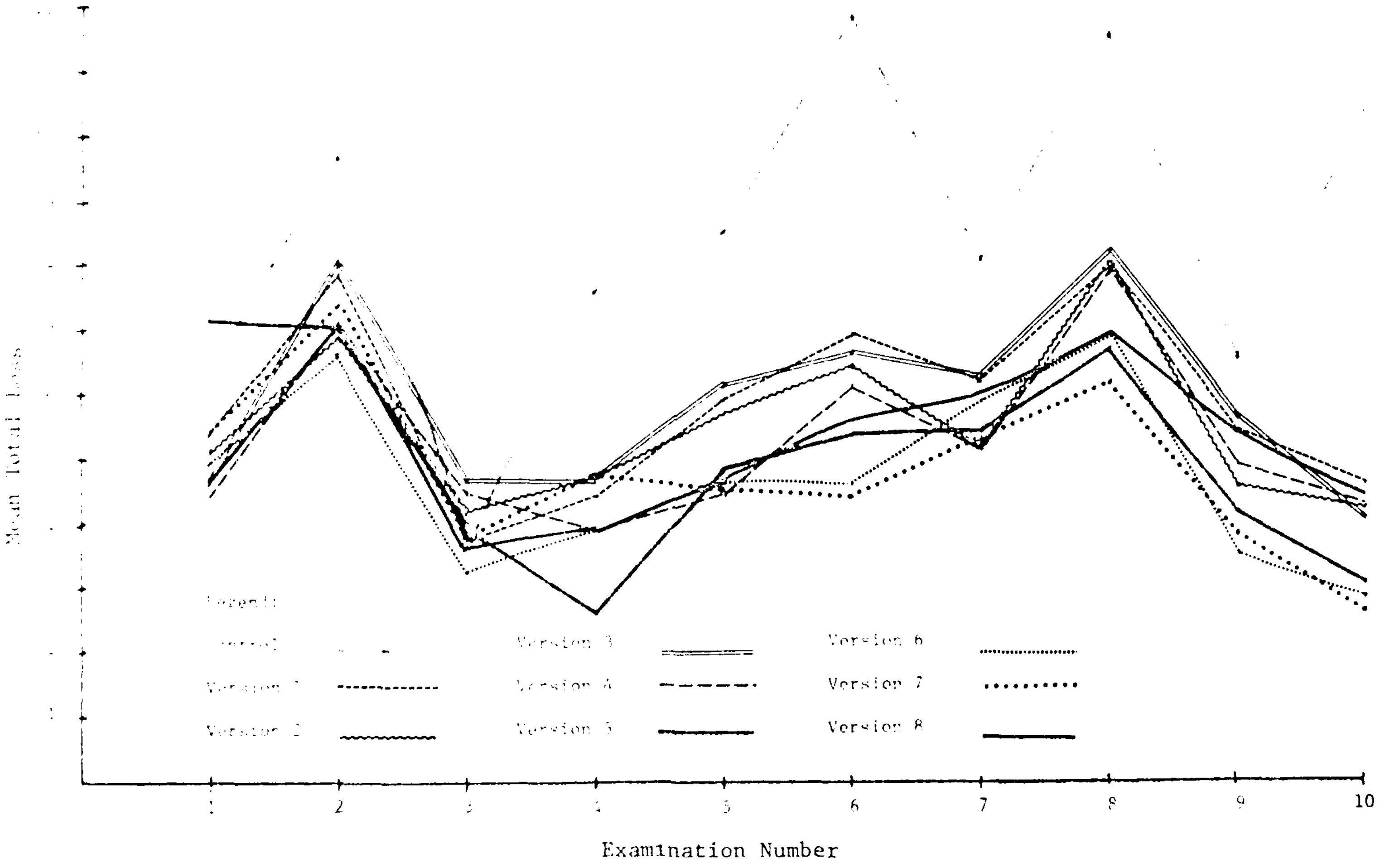


Figure 3. Testing Version by Examination Interactions for Total Loss

The Newman-Keuls test indicated no pattern of significantly different losses among examinations. Although significant differences did occur between some pairs of tests, no trend was indicated. Table 13 provides the results of this test.

The testing version by examination interaction was not significant using the conservative F-test. There was a tendency for all versions of the model to obtain approximately the same losses for each examination and to have losses less than the control except for the third examination. This is shown in Figure 3.

Total Number of Items Presented

An analysis of variance indicated significant testing version examination and testing version by examination effects ($\alpha = .05$). As for the analysis with loss as the dependent variable, a quasi-F statistic was computed. Table 14 shows the results. Tables 15 and 16 provide the descriptive statistics for the main effects.

The use of Hartley's test resulted in a rejection of the equal variance assumption. The more conservative F-tests were also significant at the .05 level.

Dunnet's test indicated that each of the adaptive testing versions required significantly ($\alpha = .05$) fewer items for presentation than the control version. A Newman-Keuls test indicated that all the adaptive tests required approximately the same number of items.

The results of the Newman-Keuls test for the examination effect are shown in Table 17. The testing version by examination interaction is displayed in Figure 4. The total number of items presented per examinee varied significantly by examination. Although the number of items per examination varied for the control version, the rank order of the means for the tests shown in Table 17 does not agree very closely with the rank order of the means for the control. Differential hierarchical configurations, objective difficulties, and hierarchical errors probably produced substantial differences in the required number of items.

Figure 4 shows only the adaptive tests results. The number of items presented per examination for the control did vary but was approximately 27 items per examination. The range in numbers of items presented for the control was [20, 34]. Hence, each mean number of items depicted in Figure 4 is substantially less than required for the control.

Figure 4 shows that there is no consistent trend for any one adaptive version to require fewer or more items than the other versions. The Newman-Keuls test for testing versions had also indicated that all the versions required the same number of items across the ten examinations.

Table 13

Newman-Keuls Tests on Losses For Examinations

| | Examination | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Item | 3 | 10 | 4 | 9 | 1 | 5 | 7 | 6 | 2 | 8 |
| Means | 3.96 | 4.40 | 4.48 | 4.86 | 5.34 | 5.60 | 5.94 | 6.45 | 7.45 | 7.80 |
| Non-Significantly Different Means[a] | + | + | + | | | | | | | |
| | | + | + | + | | | | | | |
| | | | | + | + | + | | | | |
| | | | | | + | + | + | | | |
| | | | | | | | + | + | | |
| | | | | | | | | | + | + |

Note.--α = .05.

[a] Means are non-significantly different if +'s in any one row correspond to those means.

Table 14

Analysis of Variance For Total Number of Items Presented

| Source | SS | df | MS | F |
|---|---|---|---|---|
| Between Subjects | 203293.25 | 449 | | |
| Testing Version | 186616.07 | 8 | 23327.01 | 108.2* |
| Subjects within Groups | 16677.178 | 441 | 37.82 | |
| Estimates for quasi-F calculations | | 97 | 215.60 | |
| Within Subjects | 57979.70 | 4050 | | |
| Examination | 17837.53 | 9 | 1981.95 | 292.9* |
| Testing Version by Examination | 13287.80 | 72 | 184.55 | 27.3* |
| Examination X Subjects within Groups | 26854.36 | 3969 | 6.77 | |

*p < .01.

Table 15

Descriptive Statistics of Total Number of Items Presented to Each Examinee per Testing Version

| | Total Number of Items | | |
|---|---|---|---|
| Testing Version | Mean | S. D. | Range |
| 1 | 5.56 | 5.04 | [2, 38] |
| 2 | 5.55 | 4.49 | [2, 26] |
| 3 | 6.03 | 5.28 | [2, 34] |
| 4 | 6.15 | 5.64 | [2, 33] |
| 5 | 7.65 | 5.08 | [3, 30] |
| 6 | 6.89 | 4.93 | [3, 48] |
| 7 | 7.07 | 4.96 | [3, 31] |
| 8 | 7.43 | 5.24 | [3, 49] |
| Control | 26.90 | 4.35 | [20, 34] |

Table 16

Descriptive Statistics of Total Number of Items Presented to Each Examinee per Examination

| | Total Number of Items | | |
|---|---|---|---|
| Eamination | Mean | S. D. | Range |
| 1 | 9.48 | 10.24 | [2, 34] |
| 2 | 6.69 | 8.39 | [2, 34] |
| 3 | 6.25 | 7.13 | [2, 49] |
| 4 | 7.65 | 6.15 | [2, 24] |
| 5 | 6.53 | 7.30 | [2, 24] |
| 6 | 9.74 | 7.91 | [2, 34] |
| 7 | 8.76 | 9.63 | [2, 33] |
| 8 | 12.80 | 7.70 | [2, 38] |
| 9 | 9.15 | 6.33 | [2, 26] |
| 10 | 10.98 | 7.46 | [3, 31] |

Table 17

Newman-Keuls Tests on Total Number of Items Presented for Each Examinee per Examination

| | Examination | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Item | 3 | 5 | 2 | 4 | 7 | 9 | 1 | 6 | 10 | 8 |
| Means | 6.25 | 6.53 | 6.69 | 7.65 | 8.76 | 9.15 | 9.48 | 9.74 | 10.98 | 12.80 |
| Non-Significantly Different Means [a] | + | + | | | | | | | | |
| | | + | + | | | | | | | |
| | | | | | | + | + | | | |
| | | | | | | | + | + | | |

Note. -- α = .05.

[a] Means are non-significantly different if +'s in any one row correspond to those means.

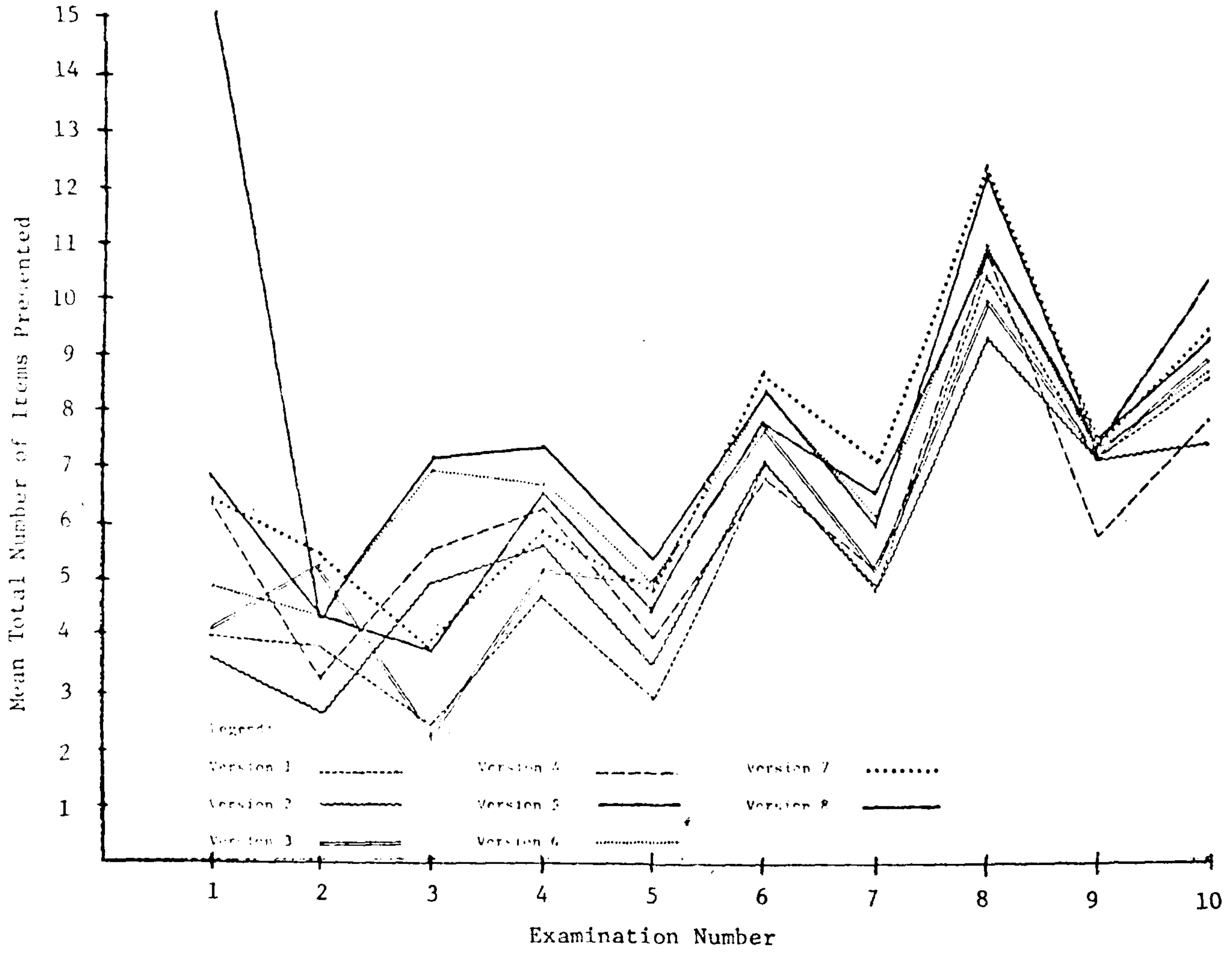


Figure 4. Adaptive Testing Version by Examination Interactions for Total Number of Items Presented.

Number of Items Presented for Item Prediction Purposes

An analysis of variance indicated significant testing version, examination, and testing version by examination effects ($\alpha = .05$). As with the other dependent variables, a quasi-*F* statistic was calculated for the testing version effect. All the effects were also significant ($\alpha = .05$) for the more conservative *F*-test, used because of the heterogeneous variances. Table 18 shows the results of the analysis. Tables 19 and 20 provide the descriptive statistics for the main effects. The results of Dunnet's test for the testing version effect showed that each adaptive test required significantly fewer ($\alpha = .05$) items than the control. The Newman-Keuls test indicated there were no significant differences among the adaptive versions.

The results of the Newman-Keuls tests for the examination effect are shown in Table 21. The testing version by examination interactions are shown in Figure 5. Although significant differences exist in numbers of items presented for the 10 examinations, Figure 5 shows that the adaptive versions vary only slightly in their relative efficiency. A version that appears to require the fewest items on examination may require the most on another test. The numl er of items required by the adaptive versions for any one test are not substantially different. The Newman-Keuls test for the testing version effect showed no adaptive test required significantly fewer items than any other version across the 10 tests.

Selection of Adaptive Testing Versions for Phase II

The intention of the Phase II study was to compare the results of some of the adaptive testing versions to those obtained in the present testing system used in the Weapons Mechanics course. The fourth and sixth adaptive testing versions were selected. The sixth version was selected because of its superior accuracy. No version was significantly superior in numbers of items presented. Solely on the basis of the mean number of items presented for item prediction, the fourth version was also selected.

The fourth version was selected on the basis of the number of items presented for item prediction purposes rather than total number of items presented. This was done because the presentation of additional items for examinees' historical data records did not prove to be adequate in most situations to use the historocal data in subsequent examinations. A change in the algorithm governing the number of additional items selected or an increase in the number of examinations would be necessary to collect sufficient bias data. Since neither alternative was feasible for Phase II, none of the additional items was presented in the Phase II study. Hence, the selection was based on a dependent variable that would feasibly be studied in the next phase, namely, the number of items presented for item prediction.

Table 18

Analysis of Variance For Number of Items Presented for Prediction Purposes

| Source | SS | df | MS | F |
|---|---|---|---|---|
| Between Subjects | 251414.84 | 449 | | |
| Testing Version | 250284.65 | 8 | 31285.58 | 256.06* |
| Subjects within Groups | 1130.19 | 441 | 2.56 | |
| Estimates for quasi-F calculations | | 72 | 122.18 | |
| Within Subjects | 18955.10 | 4050 | | |
| Examination | 2488.63 | 9 | 276.51 | 142.53* |
| Testing Version X Examination | 8752.00 | 72 | 121.56 | 62.66* |
| Examination X Subjects within Groups | 7714.47 | 3969 | 1.94 | |

*p < .01

Table 19

Descriptive Statistics of the Number of Items Presented
for Prediction Purposes for Adaptive Testing Versions

| | Number of Items | | |
|---|---|---|---|
| Testing Version | Mean | S. D. | Range |
| 1 | 2.88 | 1.50 | [2, 11] |
| 2 | 3.09 | 1.72 | [2, 9] |
| 3 | 2.92 | 1.38 | [2, 7] |
| 4 | 2.84 | 1.30 | [2, 8] |
| 5 | 3.45 | 1.60 | [2, 12] |
| 6 | 3.48 | 1.92 | [2, 14] |
| 7 | 3.47 | 1.90 | [2, 15] |
| 8 | 3.34 | 1.64 | [2, 14] |

Table 20

Descriptive Statistics of the Number of Items Presented for Prediction Purposes for Examinations

| | Number of Items | | |
|---|---|---|---|
| Examination | Mean | S. D. | Range |
| 1 | 7.13 | 9.75 | [2, 34] |
| 2 | 5.59 | 7.70 | [2, 27] |
| 3 | 4.48 | 5.53 | [2, 20] |
| 4 | 5.80 | 5.70 | [2, 21] |
| 5 | 5.33 | 6.75 | [2, 24] |
| 6 | 5.58 | 7.36 | [2, 26] |
| 7 | 6.35 | 9.55 | [2, 33] |
| 8 | 6.83 | 8.02 | [2, 29] |
| 9 | 5.23 | 7.39 | [2, 26] |
| 10 | 5.85 | 8.32 | [2, 29] |

Table 21

Newman-Keuls Tests on Number of Items Presented for Item Predictions

| | Examination | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Item | 3 | 9 | 5 | 6 | 2 | 4 | 10 | 7 | 8 | 1 |
| Means | 4.48 | 5.03 | 5.33 | 5.58 | 5.59 | 5.80 | 5.85 | 6.35 | 6.83 | 7.13 |
| Non-Significantly Different Means [a] | | + | + | | | | | | | |
| | | | | + | + | | | | | |
| | | | | | | + | + | | | |

Note.--α = .05.

[a] Means are non-significantly different if +'s in any one row correspond to those means.

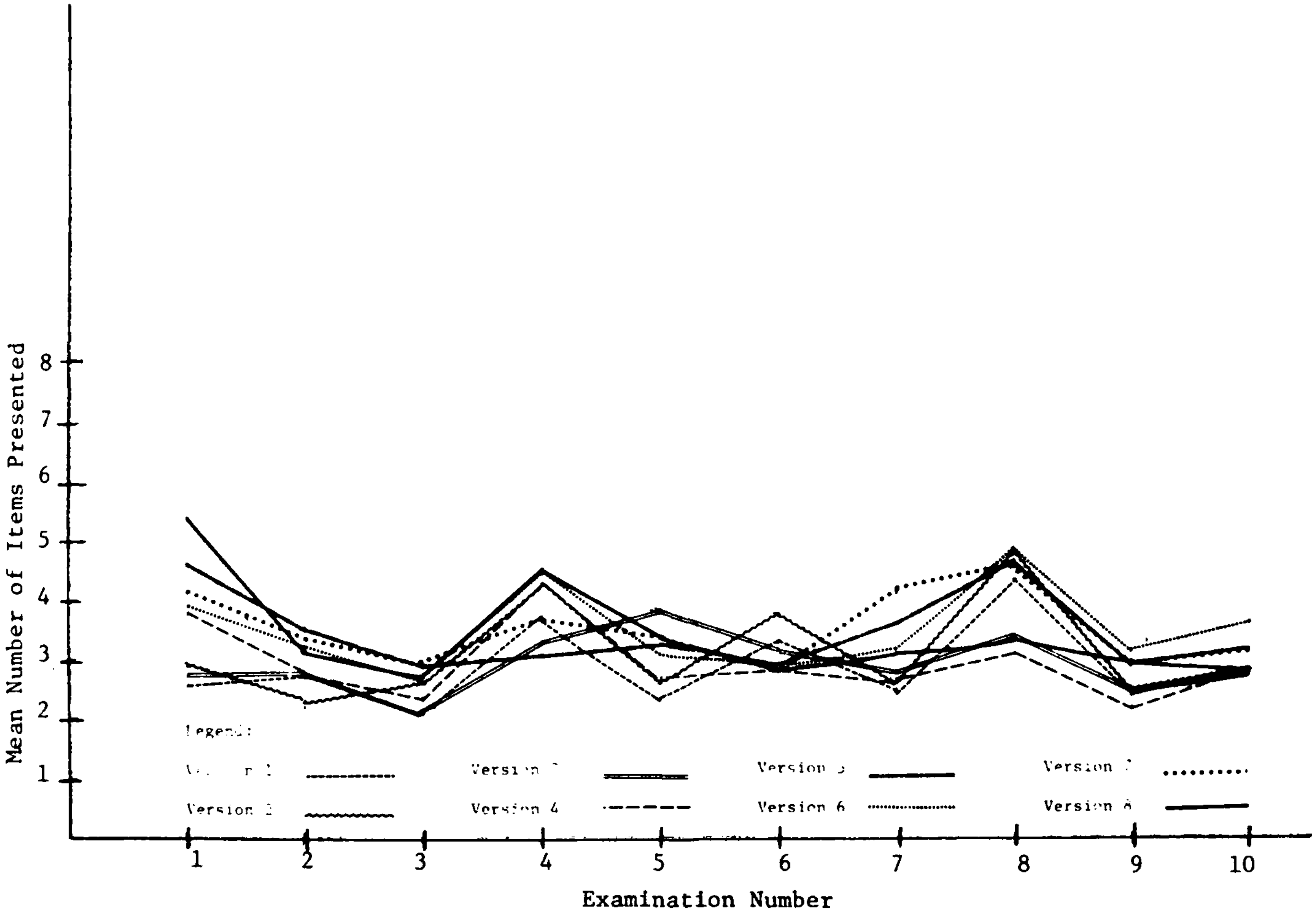


Figure 5. Testing Version by Examination Interactions for Number of Items Presented for Item Predictions

## Loss as a Function of Achievement Levels

Although the sixth version demonstrated overall superior accuracy, the losses incurred for all examinees were not the same. More importantly, the losses relative to examinees' general achievement may be small for some levels but high for others. The mean losses as a function of examinees' achievement levels are shown for the fourth and sixth versions and for the control treatment in Figure 6.

The comparison of losses with respect to achievement levels demonstrated that both adaptive testing versions performed equally well throughout the achievement range. The sixth version demonstrated a slight advantage over the fourth in the lower end of the achievement levels. This is probably due to the inconsistency check employed in the sixth version.

The adaptive testing versions had smaller losses for the middle and upper achievement levels, but this was reversed for the lower levels. This difference could be eliminated by reducing the $\alpha$-level. It may be recalled that $\beta$ was set to .1 whereas $\alpha$ was set at .2. Since the false non-mastery error would be larger than the false mastery error, a higher proportion of false classifications would be expected for those who would be at the lower achievement levels.

The adaptive testing versions may have produced more inaccurate classifications due to the paucity of data representative of poorer achieving students. Since in the data base only a small proportion of examinees did not master the objectives, the predictions made for the poorer achieving students were often based on relatively few data cases. Such was not the case for those with higher achievement levels.

## Elements of the Adaptive Testing Model

Some of the elements of the model were reviewed to see whether they appeared to be functioning adequately.

Individual Historical Records. Collection of individual examinees' historical data was not useful in this study. Although data were generally obtained for the highest categories, that is, with the highest probabilities, very few examinees had sufficient observations for the lower probability categories. The algorithm may require revision but possibly additional examinations would have overcome the problem in some categories. This area requires more study before being implemented.

Required Number of Data Base Observations. For this phase of the study 300 examinees' response vectors were used for each adaptive testing version. On the average, adaptive versions four and

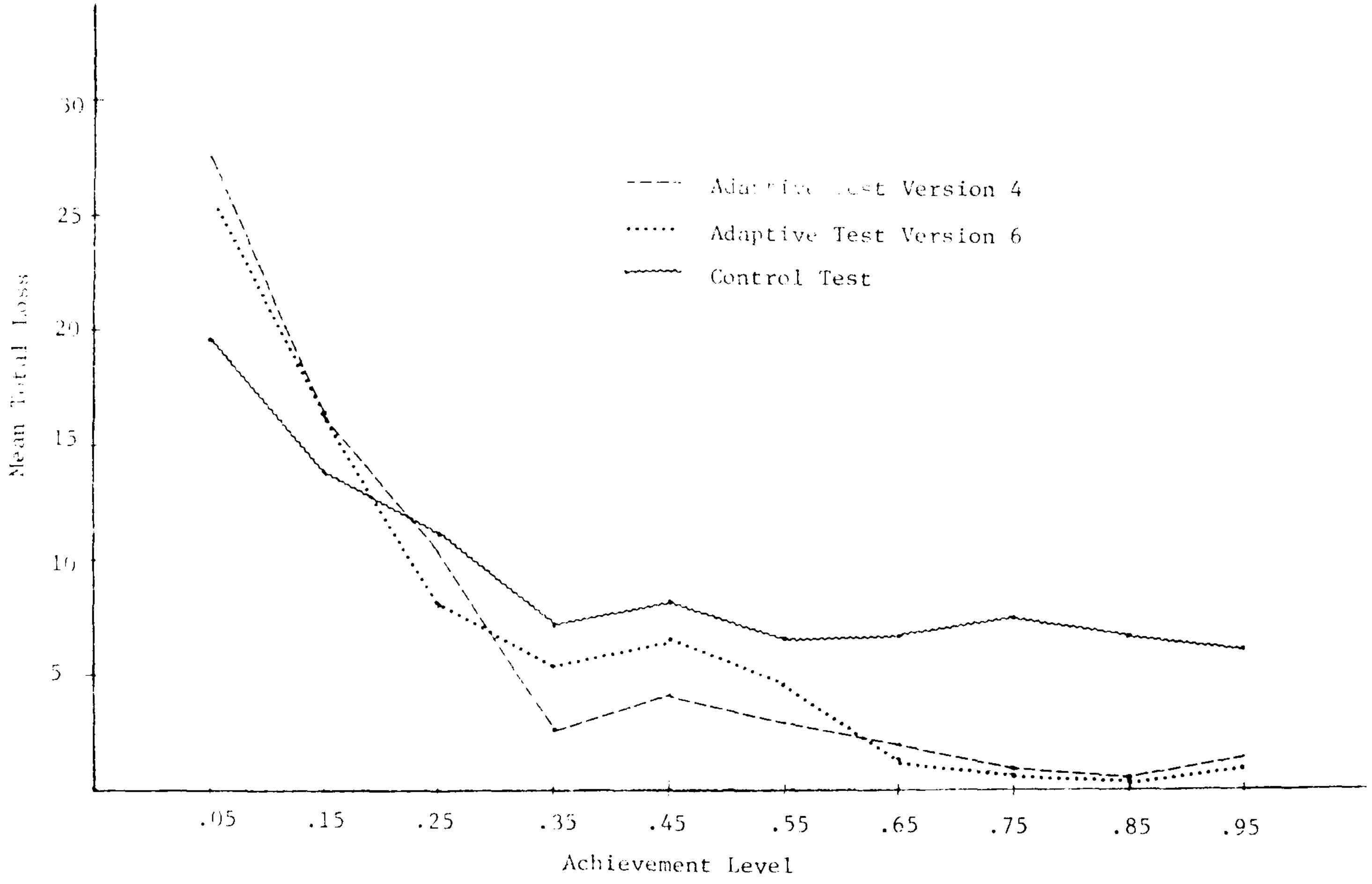


Figure 6. Mean Total Loss Corresponding to Levels of General Achievement

six had 170.15 and 152.15 matching response vectors at the conclusion of four selected examinations. For these versions, it appears that fewer response sets would be necessary in testing most examinees. But the ranges indicate that for some examinees no matching response patterns existed at the conclusion of the test. This happened for examinees at the lower achievement levels, for whom relatively little data were available. Hence it appears that fewer response sets need to be used, but more response sets reflecting the poorer performers need to be included in the data base.

## PHASE II EXPERIMENTAL DESIGN

The purpose of Phase II of the study was to make the following comparisons:

1. The relative efficiency of the fourth and sixth adaptive testing versions to each other and to the present testing method used in the Weapons Mechanics course.

2. The relative efficiency of the fourth and sixth adaptive versions to alternative adaptive testing versions requiring the first presented item to be randomly selected.

3. The classification decisions made from the adaptive tests to those made by the present method used in the Weapon Mechanics course.

Efficiency was examined in terms of the number of items presented for prediction purposes. Historical data for examinees were not collected for two reasons:

1. Phase I results indicated the historical data records appeared to be inadequate within 10 examinations. Since Phase II would have only four examinations, there would have been insufficient testing to use historical data.

2. Data obtained from the Air Force did not provide sufficient subjects taking the same forms of the examinations to match examinees across the examinations.

Hence the total number of items presented for each examinee on each adaptive test in Phase II was equivalent to the number of items presented for prediction purposes in Phase I.

The version in this phase was a testing procedure consisting of a fixed set of items for each objective. Hence all examinees answered the same set of items under the control version.

Classification decisions made by the adaptive testing versions and the control version were compared using an index defined as the number of agreements minus the number of disagreements. An agreement in classifying an examinee's performance on an objective is obtained when both indicate "non-mastery" or both indicate "mastery." Since with the adaptive testing versions performance classified as "indeterminate" dictates procedures identical to those classified as "non-mastery", an "indeterminate" classification given a true "non-mastery" state, was considered an agreement.

Alternative adaptive testing versions requiring the first presented item to be randomly selected were used as a means of varying the first item different examinees would see. If all examinees were to receive exactly the same first item, untested trainees would probably have knowledge of the item from individuals already tested. A variation of the fourth and sixth versions with randomly selected first items was used. For both the fourth and sixth versions the five items with the highest inter-item agreements were determined. From among the five items, one was randomly selected for an examinee. The response obtained was used in selecting the next item and in making other item predictions.

The $\alpha$ and $\beta$ values selected were the same as in Phase I, i.e., .2 and .1, respectively. The $\theta_0$ and $\theta_1$ values were set by using Method 3 specified in the section entitled "setting $\theta_0$ and $\theta_1$" in Appendix A.

Actual data collected on four tests of the Weapon Mechanics course were used in the computer simulations for this phase. Data were separated on the basis of examination number and form. Within each examination number by examination form category, the records were sorted in chronological order. Only data for examinations taken on or after August 11, 1977 were used. Revised examination forms implemented on that date represent the tests presently being used. Based upon the frequency of examinee records, the second form of each examination was selected for the study.

For each selected examination, from 250 to 290 response sets were available. It was not feasible to match student identification codes across the examinations since there was no control over the forms of the examinations taken by the examinees. For each examination, the first 150 response sets from among all the records sorted in ascending chronological order, were used to form the data bases. Fifty of the remaining subjects were randomly selected as the examinees who were to take the simulated adaptive tests. Hence, within each test the same 50 _S_s were used as the examinees regardless of testing version, but the same 50 _S_s were not used across examinations.

The assumed hierarchical configurations for the objectives for each block were provided by the contract monitor from the Air Force Human Resources Laboratory. The mastery score for an objective with n($\geq$ 2) items was set to n - 1, as is presently done with the conventional treatment. If n equaled one, the cutting score was set to one.

Correlated _t_-tests were used to compare adaptive testing versions. A _t_-test for a mean equal to a constant was employed for each comparison of an adaptive version to the control conversion.

## PHASE II RESULTS

All the adaptive testing procedures used in Phase II of the study demonstrated that each required significantly fewer items than the control treatment. The fourth version of the model required the presentation of fewer items than the sixth version. The modified procedures for each of these versions, namely the random selection of the first item, produced results that were the same as those obtained by their counterparts.

### Efficiency

The fourth adaptive testing version required statistically significantly (_t_ = 8.30, df = 199, p < .001) fewer items than the sixth version. The descriptive statistics for these versions (without random first item) are shown in Table 22. Although there is a statistical difference, the superior efficiency of the fourth version amounts to less than one item per examinee per test.

The adaptive testing variations involving the randomly selected first items required practically the same number of items. Table 22 shows the descriptive statistics. It should be noted that the randomly selected item would be expected to be the same as the item selected with the non-random counterpart for 20 percent of the examinees.

### Mastery/Non-Mastery Decisions

The fourth adaptive testing version had a statistically significantly (_t_ = 5.58, df = 199, p < .001) higher agreement in mastery/non-mastery classifications than the sixth version. The descriptive statistics for these versions are shown in Table 23. The average number of objectives per test was 7.25. Hence the range of the index could be [-7.25, 7.25]. A complete agreement in decisions would result .in an index value of 7.25; a complete disagreement would result in a value of -7.25. In terms of percent of agreements in decisions, the fourth and sixth versions had 92 an. 88 percent agreements with the control, respectively.

The adaptive testing variations involving the randomly selected first items had equally high indices of agreement with their counterparts. The descriptive statistics are shown in Table 23.

### Adaptive Testing Versus the Control Treatment

Separate _t_ tests were to be performed on the number of items presented for each of the four adaptive testing procedures (versions 4 and 6, with and without random first item) compared to the number required by the control. The mean number of items presented under the control treatment across the four tests was 15.25. The number of

TABLE 22

Descriptive Statistics for Phase II Treatments--
Number of Items Presented

| Statistic | Version of Adaptive Test | | | |
|---|---|---|---|---|
| | Fourth | | Sixth | |
| | Without Random First Item | With Random First Item | Without Random First Item | With Random First Item |
| Mean | 3.02 | 3.93 | 3.92 | 3.91 |
| S.D. | 1.19 | 1.34 | 1.42 | 1.43 |

TABLE 23

Descriptive Statistics for Phase II Treatments--
Index of Agreement

| Statistic | Version of Adaptive Test | | | |
|---|---|---|---|---|
| | Fourth | | Sixth | |
| | Without Random First Item | With Random First Item | Without Random First Item | With Random First Item |
| Mean | 6.15 | 6.14 | 5.54 | 5.61 |
| S.D. | 3.39 | 3.42 | 3.27 | 3.32 |

items required by the adaptive testing versions are presented in Table 24. The visual comparison of the tabled values with 15.25 reveals such large differences that no statistical test is necessary.

Since the four examinations differed in hierarchical configurations, number of objectives, and numbers of available items, Table 24 presents the percent of reduction in test items required by the adaptive testing procedures in relation to the control for each examination. The table also shows the percent of agreements in mastery/nonmastery decisions between each adaptive treatment and the control.

The results definitely show that both the fourth and the sixth adaptive testing versions, with or without the first item being selected randomly, make most of the same mastery/non-mastery decisions as are presently being made by the Air Force in its Weapons Mechanics course. But the adaptive tests make the decisions with approximately 75 percent fewer items.

## Number of Response Patterns Matching Examinees' Responses

For both phases of the study the numbers of sets of responses needed in the data bases were unknown. For the second phase it was estimated that 150 sets would be sufficient. It is desirable to have a data base with sufficiently diverse item response patterns to be able to match each examinee's response pattern in the adaptive testing situation. The results indicate that an average of 29 sets matched each examinee's set with the conclusion of each examination. The ranges in number of sets indicate that for every test and for every adaptive testing procedure no response patterns matched the examinee's pattern at the conclusion of the examination. As in Phase I, it may not be that the data base contains insufficient numbers of response patterns but that there is an insufficient number of patterns for the more poorly performing individuals. In both phases, the data bases were composed of response patterns representative in type and proportion to those patterns expected in the population of examinees. It appears that when a high proportion of examinees master the objectives, as in the Weapons Mechanics course, such a data base is insufficient for predictions of performance by non-mastering examinees. Hence, in such a situation, oversampling of non-mastering examinees may be required in order to provide adequate data for all levels of performance.

Because of the similarity of the results for all the versions in Phase I and the superior efficiency demonstrated by all the adaptive procedures in Phase II, it appears that any of the adaptive testing variations used in this study would be much more efficient than the testing procedure used by the Air Force.

Table 24

Comparison of Results of Adaptive Testing
Versions to Control Version for Each Examination

| Examination | Number of Items Presented in Conventional Version | Number of Items Presented in Adaptive Testing Version | Percent of Item Reduction | Number of Objectives | Percent of Agreements |
|---|---|---|---|---|---|
| Adaptive Testing Version 4--Without Random First Item | | | | | |
| 1 | 20 | 4.3 | 79 | 14 | 91 |
| 2 | 12 | 2.6 | 78 | 4 | 98 |
| 3 | 14 | 2.5 | 82 | 6 | 86 |
| 4 | 15 | 3.2 | 79 | 5 | 99 |
| Adaptive Testing Version 4--With Random First Item | | | | | |
| 1 | 20 | 4.1 | 80 | 14 | 91 |
| 2 | 12 | 2.6 | 78 | 4 | 98 |
| 3 | 14 | 2.0 | 86 | 6 | 86 |
| 4 | 15 | 3.0 | 80 | 5 | 99 |

(Continued on next page)

Table 24 (Cont'd)

Comparison of Results of Adaptive Testing Versions to Control Version for Each Examination

| Examination | Number of Items Presented in Conventional Version | Number of Items Presented in Adaptive Testing Version | Percent of Item Reduction | Number of Objectives | Percent of Agreements |
|---|---|---|---|---|---|
| Adaptive Testing Version 6--Without Random First Item | | | | | |
| 1 | 20 | 5.1 | 75 | 14 | 87 |
| 2 | 12 | 4.2 | 65 | 4 | 92 |
| 3 | 14 | 2.4 | 83 | 6 | 84 |
| 4 | 15 | 4.0 | 73 | 5 | 93 |
| Adaptive Testing Version 6--With Random First Item | | | | | |
| 1 | 20 | 5.0 | 75 | 14 | 88 |
| 2 | 12 | 4.2 | 65 | 4 | 92 |
| 3 | 14 | 2.4 | 83 | 6 | 84 |
| 4 | 15 | 4.0 | 73 | 5 | 93 |

## CONCLUSIONS AND RECOMMENDATIONS

Both phases of the study demonstrated conclusively that the adaptive testing model can provide practically the same results as criterion-referenced tests of fixed length for all examinees but with much greater efficiency. Simulations with data collected from trainees in the Air Force Weapons Mechanics course showed that two versions of the model reduced the number of items presented by an average of 75 percent.

All eight adaptive testing versions worked equally efficiently. The sixth version which used only item response vectors (not objective mastery/non-mastery vectors), inter-item agreement, and the inconsistency check demonstrated a slight superiority in accuracy.

The individual examinee's historical data collection procedure did not appear to work adequately. Phase I results showed that sufficient data for the higher achievement levels were obtained for some individuals but there were insufficient data for the lowest achievement levels. There were practically no data for the middle achievement levels, but this is not necessarily a problem since the most crucial levels for which to obtain historical data are the upper and lower levels. These are the levels at which item predictions are made and hence sufficient historical data at these levels would be expected to improve the accuracy of the predictions.

In Phase I, data were generated by computer; consequently, the true "mastery" or "non-mastery" state for each examinee on each objective was known. This information provided the opportunity to compare the accuracy of the adaptive testing versions to the control version. Also, because the data were generated, each examinee's general achievement level was known. In real-world situations, the true mastery/non-mastery state and the general achievement levels are not known. Hence the Phase I study provided an opportunity to investigate how accurate the adaptive testing versions were for individuals of different achievement levels. Relative to the control, the adaptive testing versions showed greater accuracy for examinees at the middle and upper achievement levels, but had poorer accuracy at the lower levels. Although the critical reason for this difference was not proved in the study, it is hypothesized that the scarcity of data representing the response patterns of more poorly performing examinees resulted in inaccurate item predictions. Hence it is recommended that at least 30 to 40 percent of the data bases from which item predictions are made be representative of the more poorly performing examinees.

One element of the model not investigated was the necessity for a valid hierarchy of objectives. In Phase I, hierarchies with small errors in hierarchical consistency were created. In Phase II unvalidated hierarchies were used. The adaptive testing versions are expected to perform more efficiently than the control for valid hierarchies. But even with the unvalidated hierarchies in

Phase II the adaptive testing versions were superior to the control. Hence the assumed hierarchies established by the Air Force may be adequate to obtain sufficient accuracy with greater efficiency for many of its training programs. In areas of criticality it may be desirable to validate hierarchies, but for many areas, the cost and time required to validate may not be necessary.

The Phase II study also showed that a minor variation in two adaptive testing versions did not alter their performance. When the first item to be presented was randomly selected from among the five items with the highest inter-item coefficients, the fourth and sixth versions were negligibly affected. This type of variation may be necessary to reduce problems with test compromise; that is, examinees knowing prior to testing which items they will obtain.

The size of the item pools for the objectives may influence accuracy and efficiency. The Phase II item pools ranged from one to six items per objective. In Phase I the item pools had 15 or 20 items per objective. Hence the data bases for the two phases were quite different with respect to the number of items on which predictions were made. Differences in the size of the item pools may affect the length of the test. Phase II showed a 75 percent reduction in the number of items presented; Phase I showed that without historical data collected (the procedure used for Phase II) the test length could be reduced by 88 percent. (See Table 24.)

## Recommendations

The results of the study indicate that the adaptive test model formulated shows such potential for extensively reducing testing time that at least one of the versions should be tested in a training environment with real trainees. Such an implementation study should exclude elements of the adaptive testing versions that are not working until additional research provides sufficient information to decide on the elements' merits. Therefore, the recommendations are separated into two areas--implementation study and research.

Implementation Study. An implementation study should be conducted to include the following:

1. Use of at least one of the adaptive testing versions or adaptations (such as with the first item being randomly selected) but without the historical data option.

2. Implementation of adaptive testing in the Weapons Mechanics course or a program with similar criticality of skills and similar instructional delivery.

3. Data bases from which predictions are made to be composed of response patterns from prior trainees but with 30 to 40 percent of each data based composed of the responses of more poorly performing trainees.

Although the adaptive testing procedures worked effectively with small numbers of items per objective, it is suggested that the different items from alternate forms of the existing tests be combined to form larger item pools for each objective. The larger item pools permit predictions to more items, with the expected benefits being better accuracy and even greater reductions in numbers of items presented than demonstrated in Phase II.

Research. Computer simulations with the use of both computer-generated and real examinee's responses should be employed to study the following issues:

1. The effects of $\alpha$ and $\beta$ on accuracy and efficiency.

2. The effects of differential proportions of response sets for examinees at differing achievement levels.

3. The accuracy and efficiency of the adaptive testing procedures relative to sizes of item pools for objectives.

4 The effects on accuracy and efficiency on different types of hierarchical configurations and different selections of the objectives of prime importance.

5. A study of the individual examinee's historical data collection and use. (How to collect sufficient data for each examinee's historical data record? Do the predictions provide greater accuracy across achievement levels and differentially between levels?)

The use of an individual examinee's historical data record may have some undesirable effects in real training environments. Although research using computer simulations may eventually demonstrate increased accuracy or the presentation of fewer items, the final resulting efficiency will be based on actual time spent in testing, not the number of items presented. The inclusion of historical data into selection and prediction of items may require more computer processing than without its use. Without the use of historical data, item selections and predictions could be predetermined for the most frequently obtained response patterns. With historical data, predetermined selection and prediction patterns would be impossible.

It should be noted that with the adaptive testing procedures, the majority of examinees tend to have response patterns matching a relatively small number of different patterns. Matching an examinee's response pattern to an existing pattern can reduce computer processing time and delay in selection and presentation of the next item.

## REFERENCES

Cohen, J. Statistical power analysis for the behavioral sciences. New York: Academic Press, 1969.

Ferguson, R. The development, implementation, and evaluation of a computer-assisted branched test for a program of individually prescribed instruction. Unpublished dissertation, University of Pittsburg, 1969.

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

Kalisch, S. J. A tailored testing model employing the beta distri bution and conditional difficulties. Journal of Computer-Based Instruction, 1974, 1, 22-28. (a)

Kalisch, S. J. The comparison of two tailored testing models and the effects of the models' variables on actual loss. Unpublished dissertation, Florida State University, 1974. (b)

Lamos, J. P. & Waters, B. K. A low-cost terminal usable for compu terized adaptive testing. Proceedings of the 1977 Computerized Adaptive Testing Conference. Arlington, VA: Personnel and Training Research Programs, Office of Navel Research, July 1978.

McCombs, B. The Air Force advanced instructional system: research and development with its adaptive component. AERA/SIGCAI Newsletter, 1977, 3(1), 7-8.

Wald, A. Sequential analysis. New York: John Wiley & Sons, 1947, reprinted by Dover Publications, 1973.

Winer, B. J. Statistical Principles in Experimental Design. New York: McGraw-Hill, 1971.

# APPENDIX A:  WALD PROBABILITY RATIO TEST

# APPENDIX A: WALD PROBABILITY RATIO TEST

The Wald sequential probability ratio test was developed by Wald (1947) as a means of making statistical decisions using as limited a sample as possible. Wald developed probability ratio tests and corresponding sequential procedures for several statistical distributions. One of the tests, that for binomial distributions, was used more recently by Ferguson (1969, 1970) for an adaptive testing application.

## Binomial Probability Ratio Test

The binomial probability ratio test was formulated by Wald 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 procedure involves the consideration of two hypotheses:

$$H_0: \ P \leq \theta_0$$

$$\text{and } H_1: \ P \geq \theta_1 \ \text{ where}$$

P is the proportion of defectives in the collection under consideration, $\theta_0$ is the maximum acceptable level of defectives to accept the collection, and $\theta_1$ is the minimum level of defectives, at or above which the collection is rejected.

The sequential aspects of the process are embodied in the capability of making one of three decisions after each element of the collection is selected. The three decisions are

1. Accept the collection
2. Reject the collection
3. Continue testing (sampling)

The decision made at any point in the process is based on the cumulative information regarding the expected proportion of defectives or non-defectives of the entire collection. Suppose that $x_1$, $x_2, \ldots x_n$ is a set of elements randomly selected without replacement from the collection under study. Suppose further that the following hypotheses are under consideration:

$$H_0: \ P \leq \theta_0$$

$$\text{and } H_1: \ P \geq \theta_1 \ , \text{ where}$$

P represents the proportion of defectives for the entire collection, and $\theta_0$ and $\theta_1$ are, respectively, the selected levels for acceptance and rejection.

Although the desirable proportion of defectives is less than or equal to $\theta_0$, an exact hypothesis $H_0'$: $P = \theta_0$ can be expected to produce results very closely approximating those obtained with the inexact hypothesis (Wald, 1947, pp. 78-79). The corresponding exact hypotheses under consideration are

$$H_0: \; P = \theta_0$$

$$H_1: \; P = \theta_1$$

The probability that the sample $x_1, x_2, \ldots x_n$ is obtained when $H_0'$ is true is expressed as

$$P_0 = f(x_1,\theta_0) \cdot f(x_2,\theta_0) \cdot \ldots f(x_n,\theta_0),$$

where f is the binomial probability function

$$f(x,\theta) = \theta^x(1-\theta)^{1-x}.$$

Similarly, the probability that the sample is obtained when $H_1$ is true is

$$P_1 = f(x_1,\theta_1) \cdot f(x_2,\theta_1) \cdot \ldots f(x_n,\theta_1).$$

Since each selected element is classified as either defective or non-defective, the value of x can be restricted to two values, specifically 0 and 1. Letting x = 0 for a non-defective, and x = 1 for a defective,

$$f(x,\theta) = \begin{matrix} \theta, \text{ for a defective} \\ 1 - \theta, \text{ for a non-defective.} \end{matrix}$$

Hence, the probability $P_0$ that the proportion of defectives equals $\theta_0$ is given by the formula

$$P_0(d,n) = \theta_0{}^d \; (1 - \theta_0)^{n-d}, \text{ where}$$

n = the number of elements in the sample
d = the number of defective elements in the sample

Similarly, $P_1(d,n) = \theta_1{}^d(1 - \theta_1)^{n-d}$.

As an example of the binomial probability ratio test, let us consider 10 randomly selected elements from a collection. Let $\theta_0 = .2$, represent the maximum proportion of defectives permitted for the entire collection. Let $\theta_1 = .4$, the proportion of defectives at or above which the collection is to be rejected. Hence the hypotheses under consideration are

$$H_0\text{:}\quad P_0 < .2 \quad \text{and}$$

$$H_1\text{:}\quad P_1 > .4$$

The corresponding exact hypotheses are

$$H_0'\text{:}\quad P_0 = .2$$

$$H_1'\text{:}\quad P_1 = .4\ .$$

Suppose three elements of the ten selected are defective, then the probability of obtaining such a sample, given that $H_0'$ is true is

$$P_0(3,10) = (.2)^3(1 - .2)^{10-3}$$

$$.0016777.$$

The probability of obtaining such a sample given that $H_1'$ is true is

$$P_1(3,10) = (.4)^3(1 - .4)^{10-3}$$

$$.0017916.$$

Having obtained probabilities for each of the two cases, one of three decisions can be made, based upon the value of the probability ratio $\frac{P_1}{P_0}$:

1. Accept the collection if $\dfrac{P_1(3,10)}{P_0(3,10)} \leq B$

2. Reject the collection if $\dfrac{P_1(3,10)}{P_0(3,10)} \geq A$

3. Continue sampling if $B < \dfrac{P_1(3,10)}{P_0(3,10)} < A$

where A and B are two positive constants ($B < A$).
Wald provides estimates for A and B by defining A and B as follows:

$$A = \frac{1 - \beta}{\alpha}$$

$$B = \frac{\beta}{1 - \alpha}$$

where $\alpha$ and $\beta$ represent two types of errors that may be made (Wald, 1947, pp. 44-48). $\alpha$ represents the probability of rejecting $H_0'$ given $H_0'$ is true, and $\beta$ represents the probability of accepting $H_0'$ when $H_1'$ is true.

The decision to reject, accept, or continue sampling is dependent upon selection of the values of $\alpha$ and $\beta$. Continuing with the example, suppose we select $\alpha = .05$ and let $\beta = .10$. Then

$$A = \frac{1 - .10}{.05} = 18$$

$$B = \frac{.10}{1 - .05} = 0.10526$$

Since $\frac{P_1(3,10)}{P_0(3,10)} = 1.07$, $B < \frac{P_1(3,10)}{P_0(3,10)} < A$. Therefore, a decision to accept or reject the collection cannot be made at the present time.

Suppose that 10 additional observations are made with no additional defective elements found. Therefore,

$$P_0(3,20) = (.2)^3 (1 - .2)^{20-3}$$
$$\approx .00018$$

$$P_1(3,20) = (.4)^3 (1 - .4)^{20-3}$$
$$\approx .0000108$$

$$\frac{P_1(3,20)}{P_0(3,20)} \approx .06$$

Since $\frac{P_1(3,20)}{P_0(3,20)} \leq B$, the collection is accepted.

In the example, two samples of 10 observations each were made, and the conclusion each time was based on the aggregate of the observations. The Wald sequential procedure permits the selection of one additional observation at a time. This latter procedure was employed by Ferguson (1969, 1970) in his adaptive testing application.

## Ferguson's Use of the Wald Procedure

Ferguson (1969, 1970) used the Wald binomial sequential method in deciding whether an examinee is a "master" or a "non-master" of an objective. An incorrect answer supplied by an examinee is analogous to a "defective" element described in the previous context of the procedure.

As an example of the procedure employed in this mastery/non-mastery context, let us arbitrarily specify the required parameters. Let $\alpha = .1$ and $\beta = .05$. Suppose that the mastery criterion for an objective is 0.8 and the non-mastery criterion is 0.6. The decisions of mastery and non-mastery can be graphically demonstrated. Figure A-1 depicts two lines ($L_0$ and $L_1$) that are the boundaries of the mastery and non-mastery regions, based upon the parameters selected. It should be noted that in order to remain consistent with the context "proportion of defectives", the mastery and non-mastery criteria must be transformed to their complements. Hence for mastery, $\theta_0 = 1 - .8$, or .2. Similarly, $\theta_1 = 1 - .6$, or .4.

The graph in Figure A-1 shows two parallel lines $L_0$ and $L_1$ that separate the region above the horizontal axis and to the right of the vertical axis into three subregions. The non-mastery region includes $L_1$ and the portion of the region above the line; the mastery region contains $L_0$ and the portion of the region below $L_0$; the no decision subregion is composed of the remainder of the region. Path $E_1$ in Figure A-1 depicts a pattern of four consecutive incorrect responses by an examinee. After each of the first three responses, no decision can be made. After the fourth, the point representing four incorrect responses for four items falls in the non-mastery zone. Path $E_2$ depicts a response pattern in which the examinee alternates with correct and incorrect responses. This process also terminates in a decision of non-mastery, but only after 12 responses have been obtained. A decision of mastery would most directly be obtained for an individual answering 11 consecutive items correctly (with no incorrect response).

The graphs of the boundaries of the mastery and non-mastery regions may be obtained by using the formulas derived by Wald (1947, p. 94). Appendix B contains a description of program WALSEQ, a FORTRAN routine which generates the intercept values and slopes for the two lines. Program WALSEQ was developed under this contract. Input to the routine requires specification of alpha, beta, mastery criterion, and the difference between the mastery and non-mastery criteria. In the last example, the criterion difference would be obtained by subtracting the non-mastery criterion, .6, from the mastery criterion, .8, resulting in a criterion difference of .2.

The values of alpha, beta, mastery criterion, and criterion difference affect the slope of the parallel lines and the vertical intercepts. Hence the selection of these parameters will also affect

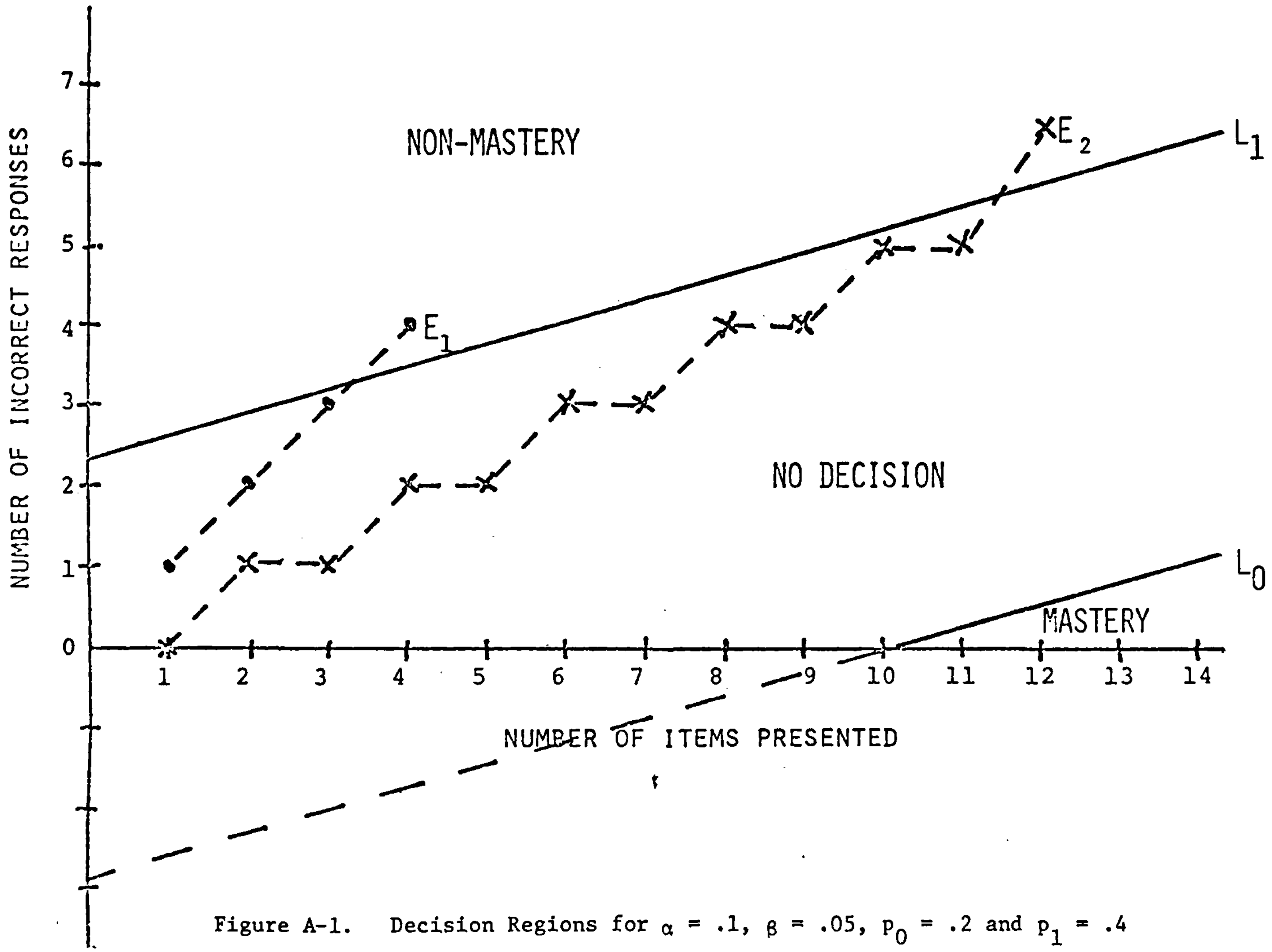


Figure A-1. Decision Regions for $\alpha$ = .1, $\beta$ = .05, $p_0$ = .2 and $p_1$ = .4

the number of items required to make a mastery/non-mastery decision. The effects of these parameters on the graphs of these lines and the expected number of items required for a decision are discussed in the next subsection.

The sequential method implies a sampling of a collection. In terms of making mastery/non-mastery decisions, this collection consists of items measuring the objective. The Wald method assumes that the elements selected are independent in the probability sense. This assumption cannot be fully met when the population is finite. The larger the population the smaller this dependence. Generally, the number of items in the population (that is, items corresponding to the same objective) is not always extremely large. Ferguson's application employed computer-generated mathematics items. Computer-generation of items is not yet possible in most instructional areas, and hence, the practical constraints of building a very large item pool for each objective forces a violation of this assumption.

The Wald binomial sequential method appears to be very satisfactory when one wishes to decide if a lot of electric light bulbs should be accepted or rejected. The light bulbs in the lot would be expected to be of the same type and size. If one were required to judge a lot of bulbs of different sizes or types, a sampling plan involving stratified selection would possibly be necessary.

Test items corresponding to the same objective may embody differences in difficulty and discrimination, and may measure different dimensions within the objective. The effects of these variables on the Wald procedure are not yet known. The sequential procedure appears to have some appealing aspects applicable to criterion-referenced testing, but also poses some questions concerning its applicability.

Mastery criterion requires a somewhat different interpretation in the context of the Wald procedure than in the context of a proportion-correct score for a small finite collection of items. In the context of the Wald procedure, a mastery criterion is the expected or desired proportion of correct responses an examinee is to provide, given infinitely many, or at least a large finite number, of items corresponding to the objective. As in the example depicted in Figure A-1 where the mastery criterion was set at 0.80, an examinee must provide correct responses to more than 80 percent of the items, when 16 or fewer items are presented. As shown in Figure A-1, the minimum number of items that may be presented in order to make a mastery decision is 11. In such a case, the examinee would not be permitted to answer any item incorrectly to be classified as a master. An examinee may incorrectly answer one item when 14 or 15 items are presented and still be given a mastery classification. In these cases, approximately 93 percent of the items must be answered correctly for a mastery classification. For presentation of 20 items the minimum acceptable percentage of correct responses drops to .90. As the number of items presented increases the minimum mastery proportion approaches .7. Hence, mastery criterion in the context of the Wald procedure is not necessarily the minimum acceptable proportion of correct responses to relatively few items. The mastery proportion may be considered an average cutting score for all subsets of items selected from an infinite collection of items.

## Effects of Differential Parameter Selection on Expected Number of Items Needed for Decision Making

In Figure A-1 a specific instance of the Wald binomial ratio test was demonstrated. In the general case, the boundaries for the mastery and non-mastery regions are parallel lines with a positive slope less than one. The slope of the lines is computed by the formula (Wald, 1947, p.94):

$$S = \frac{\log \frac{1-\theta_0}{1-\theta_1}}{\log \frac{\theta_1}{\theta_0} - \log \frac{1-\theta_1}{1-\theta_0}}$$

where $\theta_0$ is the proportion of "defectives" or items incorrectly answered to accept the collection or assume mastery; $\theta_1$ is the proportion of "defectives" or items incorrectly answered to reject the collection or assume non-mastery. Note that the slope is affected by $\theta_0$ and $\theta_1$ but not by $\alpha$ or $\beta$. Recall that $\theta_0$ = (1 - mastery criterion).

The intercepts of the lines for the mastery and non-mastery boundaries are computed from the following formulas, respectively (Wald, 1947, p.94):

$$h_0 = \frac{\log \frac{\beta}{1-\alpha}}{\log \frac{\theta_1}{\theta_0} - \log \frac{1-\theta_1}{1-\theta_0}}$$

$$h_1 = \frac{\log \frac{1-\beta}{\alpha}}{\log \frac{\theta_1}{\theta_0} - \log \frac{1-\theta_1}{1-\theta_0}}$$

The intercepts are affected by $\theta_0$, $\theta_1$, $\alpha$, and $\beta$.

The effects upon the slope and intercepts by changes in $\alpha$, $\beta$, mastery criterion $(1 - \theta_0)$, and criterion difference (mastery criterion minus non-mastery criterion), that is $(1 - \theta_0) - (1 - \theta_1)$ or $\theta_1 - \theta_0$ are presented in Table A-1. The concept of a criterion difference is analogous to the concept of effect size in statistically testing the difference between the means of two groups. In such statistical testing, when $\alpha$ and $\beta$ remain constant, the number of observations required to detect a significant difference may be reduced as the anticipated effect size increases (see Cohen, 1969).

Although Table A-1 summarizes the effects of $\alpha$, $\beta$, mastery criterion, and criterion difference on the slope and intercepts, the ramifications in terms of required numbers of items for decision making are not always apparent. Therefore, Table A-2 provides a summary of the effects of the variables on minimum numbers of items to be presented for mastery and for non-mastery decisions, and the effects on the expected number of items required for each decision. In Table A-2, the minimum number of items necessary for mastery/non-mastery decisions refers to cases in which a subject correctly answers all items or incorrectly answers all the items presented. Such response patterns produce decisions with the presentation of the fewest items.

## Expected Number of Items Needed to Make Mastery/Non-Mastery Decisions

An expected number of observations necessary to reach classification decisions given the true value of the proportions of defectives in the collection has been addressed by Wald (1973, pp. 99-101). Figure A-2 is an example of a typical average sample number function. In the figure, n represents the number of observations required, p is the proportion of defectives in the lot. The expected value of n, $E_p(n)$, generally increases as the proportion of defectives increases from zero through $p_0$, the proportion of defectives for acceptance of the lot. A change from increasing values of $E_p(n)$ occurs between $p_0$ and $p_1$, with decreasing values of $E_p(n)$ for values of p greater than $p_1$.

Table A-1

Effects on Slope and Vertical Intercepts of Mastery and Non-Mastery Boundaries when Parameters are Varied

| Conditions | | Effect on* | | |
|---|---|---|---|---|
| | | | INTERCEPTS | |
| Parameters Fixed | Independent Variable* | Slope | Mastery | Non-Mastery |
| Mastery Criterion | Alpha | None | Increases | Decreases |
| Criterion Difference | | | (Effect on non-mastery intercept greater than on mastery intercept | |
| Beta | | | | |
| Mastery Criterion | Beta | None | Increases | Decreases |
| Criterion Difference | | | (Effect on mastery intercept greater than on non-mastery intercept | |
| Alpha | | | | |
| Criterion Difference | Mastery Criterion | Decreases | Decreases | Increases |
| Alpha | | (Approaching zero) | | |
| Beta | | | | |
| Mastery Criterion | Criterion Difference | Increases | Increases | Decreases |
| Alpha | | | (Both intercepts approaching zero) | |
| Beta | | | | |

*Effects on slope and intercepts are expressed in relation to increasing values of the independent variables.

Table A-2
Effects on Minimum Number of Items and Expected Numbers of Items Necessary for Making Mastery and Non-Mastery Decisions

| Conditions | | Effects on * | | | |
|---|---|---|---|---|---|
| | Independent | Minimum Number of Items | | Range of Expected Number of Items | |
| Parameters Fixed | Variable * | Mastery | Non-Mastery | Mastery | Non-Mastery |
| Mastery Criterion<br>Criterion Difference<br>Beta | Alpha | Very little change; if any, decrease | Decrease | Decrease<br>(Decrease is greater for non-mastery) | Decrease |
| Mastery Criterion<br>Criterion Difference<br>Alpha | Beta | Decrease | Very little change; if any, decrease | Decrease<br>(Decrease is greater for mastery) | Decrease |
| Criterion Difference<br>Alpha<br>Beta | Mastery Criterion | Increase | Decrease | Generally increase to a point and then decrease | Generally increase to a point and then decrease |
| Mastery Criterion<br>Alpha<br>Beta | Criterion Difference | Decrease | Decrease | Generally decrease | Generally decrease |

*Effects on minimum and expected numbers of items are expressed in relation to increasing vaues of the independent variables.

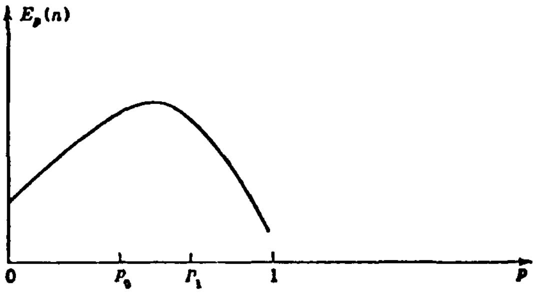


Fig. A-2. Typical Average Sample Number Function (Abraham Wald, Sequential Analysis, Dover Publications, Inc. Reprinted by permission.)

Formulas for determining $E_p(n)$ are provided in Wald's publication. The values of $E_p(n)$ are dependent upon $\alpha$, $\beta$, $p_0$, and $p_1$ values. The ramifications of the values of these variables upon the expected required number of items are discussed in ensuing sections.

The formulas for estimating $E_p(n)$ were used in program WALSEQ. A description of the program appears in Appendix B.

Setting $\alpha$, $\beta$,

The expected number of items needed to make mastery/non-mastery classifications are affected by $\alpha$ and $\beta$. The lower the values of $\alpha$ and $\beta$, the higher the expected number of items needed.

The selection of values for $\alpha$ and $\beta$ should be made on the basis of the importance of accurately classifying examinees. The effect of specifying small $\alpha$ and $\beta$ may require the generation of many items, or items with extremely high discriminations. The effort and cost involved in producing highly discriminating items for less crucial objectives may not be justified. Hence the selection of values for $\alpha$ and $\beta$ should be based on the criticality of accurate classification for each individual objective.

A suggested method of selecting values for $\alpha$ and $\beta$ is based upon the criticality levels for specified factors related to the two potential errors. The Type I error ($\alpha$) is the risk in classifying an individual as "non-master" when actually the individual is a "master." The Type II error ($\beta$) is the risk in classifying an individual as a "master" when actually the individual is a "non-master." The method requires a determination of the criticality (low, medium, high) for each factor j, selection of an $\alpha_j$ value from a given range, and selection of equal to the minimum $\alpha_j$ value. The same method is employed to establish $\alpha$ value for $\beta$.

In this report, the factors and values for $\alpha$ and $\beta$ corresponding to levels of those factors are only examples. The factors and values have neither been validated nor obtained by a consensus of experts. The factors and values are provided as a guide for further exploration and research.

For the Type I error, the following three factors concerning the impact of the level on potentially incorrect decisions have been identified:

1. Instruction--required training resources (personnel and materials) to provide additional training.

2. Trainee Attitudes--the attitude of trainees when assigned instruction on objectives that have been mastered; trainee frustration; and corresponding impact on performance in the remaining portion of the program and on the job.

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

The factors are not necessarily mutually exclusive, although no two are intended to incorporate identically the same elements. For example, suppose that a trainee has been incorrectly classified as a non-master on an objective. The corresponding instruction may require a moderate amount of time and extensive use of instructional resources. Hence the criticality of the instructional resources factor would be classified as "high." The cost/time factor would possibly be classified as "medium."

Table A-3 displays a sample matrix of suggested ranges of values for each factor by criticality level. (The values in the matrix are for illustrative purposes only.) The matrix would be used as an aid in selecting an $\alpha_j$ value for each of the j factors. It is noted that the criticality levels are not necessarily the same for each factor. Also, the same $\alpha$ value may appear at different levels for different factors. The potential impact of a wrong decision on an objective that comes at the conclusion of a long, expensive training session could be very high. A very low probability of false non-mastery classification would probably be desirable--possibly as low as .001. For other factors, it might never be practical or sufficiently critical to select an $\alpha_j$ value less than .01.

In order for the matrix in Table A-3 to be used objectively, measurable criteria would necessarily be assigned to each criticality level for each factor.

Table A-3

Sample Matrix of Alpha Values Related To a Type I Error (Classifying a True Master as a Non-Master)

| | FACTORS | | |
|---|---|---|---|
| Criticality Level | Instruction | Attitude | Cost/Time |
| | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ |
| Low | .1 - .2 | .2 - .3 | .1 - .3 |
| Medium | .06 - .09 | .08 - .1 | .02 - .09 |
| High | .01 - .05 | .01 - .07 | .001 - .01 |

$\alpha$ = minimum of ($\alpha_1$, $\alpha_2$, $\alpha_3$)

Note: The values supplied in the matrix are for illustrative purposes only.

For example, for the "impact on instruction" factor, the following measurable criteria might be employed:

| Criticality Level | Amount of Time Required for Corresponding Instruction |
|---|---|
| Low | Two minutes or less |
| Medium | More than 2 but less than 10 minutes |
| High | 10 or more minutes |

The same type of process would be used to select a value for the Type II error ($\beta$). The following four factors could be considered in selecting $\beta$:

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

2. Prerequisite in Instruction--potential ramifications upon future instruction (in the present training program or thereafter), especially if the skill is prerequisite to many other tasks.

3. Time/Cost--potential loss or destruction of equipment (either in training or on the job).

4. Trainee's View of the Training--potential negative view by trainee when classified as master, although trainee feels he is not a master; negative view of trainee in the on-the-job situations when previous training does not appear to have sufficiently prepared him for the job.

## Setting $\theta_0$ and $\theta_1$

In the context of "defectives" $\theta_0$ is the maximum acceptable level of defectives to accept the collection. In terms of "mastery" of an objective, the proportion of items answered correctly is used. Therefore, let $q_0$ represent the mastery level. Then $q_0$ is equivalent to $1 - \theta_0$. Similarly, since $\theta_1$, represents the minimum level of defectives at or above which a collection is rejected, $q_1$ will be used in terms of proportion correct and is equivalent to $1 - \theta_1$. Hence, when the mastery and non-mastery values $q_0$ and $q_1$ have been selected, $\theta_0$ and $\theta_1$ are readily obtained.

Three methods for selecting mastery and non-mastery criteria are outlined below:

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 $q_1$.

AD-A081 855 CONTROL DATA EDUCATION CO MINNEAPOLIS MN F/G 5/9
COMPUTERIZED INSTRUCTIONAL ADAPTIVE TESTING MODEL: FORMULATION --ETC(U)
FEB 80 S J KALISCH F33615-77-C-0071

UNCLASSIFIED AFHRL-TR-79-33 NL

2 of 2
AD
A081855

END
DATE
FILMED
4-80
DTIC

## The Effects of Item and Objective Discrimination

The discriminating power of the items affects the expected number of items required to make mastery/non-mastery decisions. If the items provide little discrimination, the number of items necessary for decision making may be increased beyond practical cost and developmental time limits. With respect to the three methods of setting $\theta_1$, and $\theta_2$, Methods 1 and 3 provide estimates of the discrimination of the objectives, whereas Method 2 does not.

Objective discrimination is defined as $\overline{q}_0 - \overline{q}_1$, where $\overline{q}_0$ and $\overline{q}_1$, are the mean proportion of items answered correctly by masters and non-masters of the objective, respectively. Item discrimination may be similarly defined as $r_0 - r_1$, where $r_0$ and $r_1$, are the respective proportions of masters and non-masters of the objective who have answered the item correctly. An estimate for the objective discrimination for n corresponding items would be the average of the n item discriminations.

Objective discrimination is directly related to the criterion difference discussed in the context of the Wald binomial probability ratio test. The criterion difference may be represented as $\theta_1 - \theta_0$. The criterion difference $\theta_1 - \theta_0$ equals $q_0 - q_1$, the objective discrimination.

An increase in criterion difference, and hence discrimination, decreases the expected number of items required to make mastery/non-mastery decisions. This is demonstrated in Tables A-4 and A-5. For fixed $\alpha$ and $\beta$, the expected number of items necessary for both classifications is reduced and the range in expected number of items is also diminished. The tables indicate that for each given $\alpha$, $\beta$, and criterion difference the number of items expected to make a non-mastery classification is greater than the number required to make a mastery decision, although for more highly discriminating items the expected numbers for each decision are nearly equal. Although Tables A-4 and A-5 specify values for the one case in which $\theta_0 = .2$, i.e., mastery criterion equals .8, the same pattern exists for other values of $\theta_0$, relatively close to $\theta_0 = .2$ (e.g., .1, .3, .4).

The level of item discrimination has affects upon test development costs and testing time for examinees. Objectives with items of low discrimination will require large pools of items to be developed. Writing items with higher discriminations may require more development time and dollars per item, but will require fewer items to be developed. In order to minimize testing time for examinees, it would be desirable to present as few items as possible without jeopardizing the ability to make accurate classifications. The desirability of reducing testing time and the effects upon development costs suggest the development of items with sufficiently high discriminations.

Table A-4

Ranges of Expected Numbers of Items Needed to Make a Mastery Decision When $\theta_0 = .2$

| Value of | | Criterion Difference | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $\alpha$ | $\beta$ | .1 | .2 | .3 | .4 | .5 | .6 | .7 |
| .01 | .01 | 35, 176 | 16, 50 | 10, 24 | 7, 14 | 5, 9 | 4, 6 | 3, 4 |
| .01 | .05 | 23, 114 | 11, 32 | 7, 16 | 5, 9 | 4, 6 | 3, 4 | 2, 3 |
| .05 | .01 | 35, 163 | 16, 46 | 10, 22 | 7, 13 | 5, 8 | 4, 6 | 3, 4 |
| .05 | .05 | 23, 103 | 11, 29 | 7, 14 | 5, 8 | 4, 5 | 3, 4 | 2, 2 |
| .05 | .10 | 17, 78 | 8, 22 | 5, 11 | 4, 6 | 3, 4 | 2, 3 | 2, 2 |
| .10 | .05 | 22, 93 | 11, 26 | 7, 13 | 5, 8 | 3, 5 | 3, 3 | 2, 2 |
| .10 | .10 | 17, 69 | 8, 20 | 5, 10 | 4, 6 | 3, 4 | 2, 3 | 2, 2 |
| .10 | .15 | 14, 55 | 7, 16 | 5, 7 | 3, 5 | 2, 3 | 2, 2 | 1, 2 |
| .10 | .20 | 12, 45 | 6, 13 | 4, 6 | 3, 4 | 2, 3 | 2, 2 | 1, 1 |
| .15 | .10 | 17, 61 | 8, 17 | 5, 9 | 4, 5 | 3, 3 | 2, 2 | 2, 2 |
| .15 | .15 | 13, 48 | 7, 14 | 4, 7 | 3, 4 | 2, 3 | 2, 2 | 1, 1 |
| .15 | .20 | 11, 39 | 6, 11 | 4, 6 | 3, 3 | 2, 2 | 2, 2 | 1, 1 |
| .20 | .10 | 1, 53 | 8, 15 | 5, 8 | 4, 5 | 3, 3 | 2, 2 | 1, 1 |
| .20 | .15 | 13, 41 | 6, 12 | 4, 6 | 3, 4 | 2, 2 | 2, 2 | 1, 1 |
| .20 | .20 | 11, 33 | 5, 10 | 3, 5 | 3, 3 | 2, 2 | 2, 2 | 1, 1 |

Table A-5

RANGES OF EXPECTED NUMBERS OF ITEMS NEEDED TO MAKE A NON-MASTERY DECISION WHEN $\theta_0 = .2$

| Value of |  | Criterion Difference |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|
| $\alpha$ | $\beta$ | .1 | .2 | .3 | .4 | .5 | .6 | .7 |
| .01 | .01 | 12, 159 | 7, 43 | 6, 20 | 5, 11 | 4, 7 | 4, 5 | 3, 4 |
| .01 | .05 | 12, 148 | 7, 39 | 5, 18 | 5, 10 | 4, 7 | 4, 5 | 3, 4 |
| .05 | .01 | 8, 103 | 5, 27 | 4, 13 | 3, 7 | 3, 4 | 3, 3 | 2, 2 |
| .05 | .05 | 8, 94 | 5, 25 | 4, 11 | 3, 6 | 3, 4 | 3, 3 | 2, 2 |
| .05 | .10 | 8, 84 | 5, 22 | 4, 10 | 3, 6 | 3, 4 | 2, 3 | 2, 2 |
| .10 | .05 | 6, 70 | 4, 19 | 3, 8 | 3, 5 | 2, 3 | 2, 2 | 1, 2 |
| .10 | .10 | 6, 62 | 4, 16 | 3, 7 | 3, 4 | 2, 3 | 2, 2 | 1, 2 |
| .10 | .15 | 6, 55 | 4, 14 | 2, 6 | 2, 4 | 2, 2 | 1, 2 | 1, 2 |
| .10 | .20 | 6, 48 | 4, 13 | 3, 6 | 2, 3 | 2, 2 | 1, 2 | 1, 2 |
| .15 | .10 | 5, 49 | 3, 13 | 2, 6 | 2, 3 | 2, 2 | 1, 2 | 1, 2 |
| .15 | .15 | 5, 43 | 3, 11 | 2, 5 | 2, 3 | 2, 2 | 1, 2 | 1, 2 |
| .15 | .20 | 5, 37 | 3, 10 | 2, 4 | 2, 2 | 1, 2 | 1, 2 | 1, 2 |
| .20 | .10 | 4, 40 | 3, 10 | 2, 5 | 2, 3 | 1, 2 | 1, 2 | 1, 2 |
| .20 | .15 | 4, 34 | 3, 9 | 2, 4 | 2,2 | 1, 2 | 1, 2 | 1, 1 |
| .20 | .20 | 4, 29 | 3, 7 | 2, 3 | 2, 2 | 1, 2 | 1, 2 | 1, 1 |

Estimates of item and objective discrimination are dependent upon the sample upon which the estimates are based. Hence, adequate representation of both masters and non-masters of an objective would be necessary to obtain accurate estimates of discrimination.

## APPENDIX B: A Description of Program WALSEQ

Program WALSEQ is a FORTRAN routine developed under this contract, providing the following results related to the Wald binomial probability ratio test:

1. Slope and intercepts for mastery/ non-mastery region boundaries. (See Figure A-1.)
2. Minimum and expected numbers of items for mastery/non-mastery classification.

Examples of the printed output are shown on the following two pages.

## WALD SEQUENTIAL METHOD
## NUMBERS OF ITEMS FOR CLASSIFICATION DECISIONS

ALPHA = .2000 BETA = .1000 CRITERION DIFFERENCE = .2500

| | MINIMUM NUMBER OF ITEMS FOR DECISION | | EXPECTED NUMBER OF ITEMS NEEDED FOR CLASSIFICATION | | |
|---|---|---|---|---|---|
| MASTERY CRITERION | MASTERY | NON-MASTERY | MASTERY | NON-MASTERY | NEITHER |
| .8000 | 6 | 2 | [ 6, 10] | [ 2, 7] | [ 7, 11] |
| .8100 | 6 | 2 | [ 6, 10] | [ 2, 7] | [ 7, 11] |
| .8200 | 6 | 2 | [ 6, 10] | [ 2, 6] | [ 6, 10] |
| .8300 | 6 | 2 | [ 6, 10] | [ 2, 6] | [ 6, 10] |
| .8400 | 6 | 2 | [ 6, 10] | [ 2, 6] | [ 6, 10] |
| .8500 | 5 | 2 | [ 6, 10] | [ 2, 6] | [ 6, 10] |
| .8600 | 7 | 2 | [ 7, 9] | [ 2, 6] | [ 6, 9] |
| .8700 | 7 | 2 | [ 7, 9] | [ 2, 5] | [ 5, 9] |
| .8800 | 7 | 2 | [ 7, 9] | [ 2, 5] | [ 5, 9] |
| .8900 | 7 | 2 | [ 7, 9] | [ 2, 5] | [ 5, 9] |
| .9000 | 7 | 2 | [ 7, 9] | [ 2, 5] | [ 5, 9] |
| .9100 | 7 | 2 | [ 7, 8] | [ 2, 4] | [ 4, 8] |
| .9200 | 7 | 2 | [ 7, 8] | [ 2, 4] | [ 4, 8] |
| .9300 | 7 | 1 | [ 7, 8] | [ 1. 4] | [ 4, 8] |
| .9400 | 7 | 1 | [ 7, 8] | [ 1, 3] | [ 3, 8] |
| .9500 | 7 | 1 | [ 7, 7] | [ 1, 3] | [ 3, 7] |
| .9600 | 7 | 1 | [ 7, 7] | [ 1, 3] | [ 3, 7] |
| .9700 | 7 | 1 | [ 7, 7] | [ 1, 2] | [ 2, 7] |
| .9800 | 8 | 1 | [ 6, 8] | [ 1, 2] | [ 2, 6] |
| .9900 | 8 | 1 | [ 6, 8] | [ 1, 1] | [ 1, 6] |

## WALD SEQUENTIAL METHOD
## PARAMETERS FOR MASTERY/NON-MASTERY BOUNDARIES

**ALPHA = .2000 BETA = .1000 CRITERION DIFFERENCE = .2500**

| | | *** | INTERCEPTS | | *** |
|---|---|---|---|---|---|
| | | --- MASTERY | --- | --- NON-MASTERY | --- |
| MASTERY CRITERION | SLOPE | VERTICAL | HORIZONTAL | VERTICAL | HORIZONTAL |
| .8000 | .3160 | -1.7539 | 5.5497 | 1.286 | -4.0142 |
| .8100 | .3053 | -1.7202 | 5.6339 | 1.2442 | -4.0750 |
| .8200 | .2946 | -1.6844 | 5.7180 | 1.2184 | -4.1359 |
| .8300 | .2838 | -1.6466 | 5.8021 | 1.1910 | -4.1967 |
| .8400 | .2730 | -1.6067 | 5.8861 | 1.1621 | -4.2575 |
| .8500 | .2621 | -1.5645 | 5.9701 | 1.1316 | -4.3183 |
| .8600 | .2511 | -1.5201 | 6.0542 | 1.0995 | -4.3790 |
| .8700 | .2400 | -1.4733 | 6.1381 | 1.0657 | -4.4398 |
| .8800 | .2289 | -1.4241 | 6.2221 | 1.0300 | -4.5005 |
| .8900 | .2176 | -1.3722 | 6,3061 | .9925 | -4.5612 |
| .9000 | .2062 | -1.3176 | 6.3900 | .9530 | -4.6219 |
| .9100 | .1946 | -1.2600 | 6.4739 | .9114 | -4.6826 |
| .9200 | .1829 | -1.1991 | 6.5578 | .8673 | -4.7433 |
| .9300 | .1708 | -1.1345 | 6.6416 | .8206 | -4.8040 |
| .9400 | .1584 | -1. 0656 | 6.7255 | .7708 | -4.8646 |
| .9500 | .1456 | -.9916 | 6.8093 | .7172 | -4.9252 |
| .9600 | .1322 | -.9110 | 6.8931 | .6589 | -4.9859 |
| .9700 | .1177 | -.8214 | 6.9769 | .5941 | -5.0465 |
| .9800 | .1017 | -.7177 | 7.0607 | .5191 | -5.1071 |
| .9900 | .0820 | -.5859 | 7.1445 | .4238 | -5.1677 |

## APPENDIX C: Directions for Supplying Decision Loss Values

Adaptive testing may be described as a procedure in which an examinee is presented with test items that are appropriate for his level of performance. By presenting only those items that are appropriate to the individual's performance levels, an adaptive testing procedure can provide a more accurate and more efficient system of ascertaining mastery or non-mastery of instructional objectives.

### The Problem

Suppose you are a "decision maker" in an instructional assessment program. You are planning to implement an adaptive testing procedure. The testing process requires the use of certain parameters. You as "decision maker" are to supply these values.

### Parameters in the Adaptive Testing Procedure

Suppose that the instructional objectives in each block of instruction are classified as "of prime concern" and "of secondary concern." The objectives of prime concern are those objectives which are of most importance and are generally the terminal objectives of the block. The objectives of secondary concern are generally the enabling or prerequisite objectives--those that serve as a means to achieving the terminal objectives.

The parameters you are to specify indicate the relative losses associated with various classification decisions for the two types of objectives. For example, you would consider it undesirable to classify an individual's performance on an objective as "non-mastery" when in actuality the skill had been mastered. As undesirable as this incorrect classification would be, you might not consider it as undesirable as classifying an individual's performance as "mastery" when in actuality the skill had not been mastered. Although with each of these incorrect decisions, losses with regard to accuracy may be associated, the losses may not be equal. Similarly, greater losses in inaccurate decisions may be associated with objectives of prime concern than with objectives of secondary concern.

Your decision-making task is to supply the parameters that describe the relative losses for various correct and incorrect mastery/non-mastery classifications. The 12 parameters you are to provide are represented by the $L_i$'s ($i = 1, 12$) in Tables C-1 and C-2.

Table C-1

Matrix of Loss Values Associated
with Objectives of Primary Concern.

| Classification Decision | True Classification | |
|---|---|---|
| | Mastery | Non-Mastery |
| Mastery | $L_1$ | $L_2$ |
| Non-Mastery | $L_3$ | $L_4$ |
| Indeterminable | $L_5$ | $L_6$ |

Table C-2

Matrix of Loss Values Associated
with Objectives of Secondary Concern

| Classification Decision | True Classification | |
|---|---|---|
| | Mastery | Non-Mastery |
| Mastery | $L_7$ | $L_8$ |
| Non-Mastery | $L_9$ | $L_{10}$ |
| Indeterminable | $L_{11}$ | $L_{12}$ |

The values you provide for the $L_i$'s are to reflect the relative losses associated with accurate/inaccurate decisions. The values may incorporate such factors as (a) increased instructional costs, (b) amounts of time for instruction, and (c) effects on the trainee's morale. Inaccurate classification may result in the unnecessary assignment of instructional remediation. Likewise, other inaccurate classifications may result in a trainee's being placed on the job with inadequate or unmastered skills.

The loss values you provide reflect "relative" losses. The values are not absolute losses. For example, if in Table C-1, $L_2$ equals 10 and $L_3$ equals 5, the loss ($L_2$) associated with classifying as "mastery" performance that is actually "non-mastery" is twice as great as the loss ($L_3$) associated with a "non-mastery" decision when the true situation is "mastery." These same relative losses could have been specified as $L_2 = 2$ and $L_3 = 1$. One loss would still be twice that of the other. Of course, in providing 12 relative loss values, one must in essence consider the relation of all pairs of loss values.

The loss values to be provided are to be nonnegative integers (0, 1, 2, ...). Losses associated with correct decisions may be zero since no loss in accuracy would be associated with a correct decision.

Due to insufficient data collection, it may not be possible to classify performance as "mastery" or "non-mastery." In such cases, performance is to be categorized as "indeterminable", and assumed to be "non-mastery." Hence, due to an examinee's set of responses to test items or an insufficient number of items presented, the trainee is assumed not to have demonstrated performance classifiable as "mastery" and is assumed not to have mastered the skills. Because you may consider losses associated with the "indeterminable" classification to be different than those for a "non-mastery" decision, you are given the opportunity to make the distinction.

Assumptions

In specifying the loss values, make the following additional assumptions:

1. Since incorrect decisions result in greater losses in accuracy than correct decisions, $L_2$ and $L_3$ should be greater than $L_1$ and $L_4$. Likewise $L_8$ abd $L_9$ should also be greater than $L_7$ abd $L_{10}$.

2. Incorrect decisions for objectives of primary concern should be greater than for objectives of secondary concern. Hence, $L_2$ and $L_3$ should be greater than $L_8$ and $L_9$, respectively. Likewise it is expected that $L_5$ is to be greater than $L_{11}$.

3. The loss values to be provided represent the losses related to the possible decisions that may be made for each objective.

Your Decisions

In relation to the situation and assumptions discussed, you are asked to supply whole number loss values (0, 1, 2, 3, ...) for the 12 $L_i$ values in Tables C-1 and C-2. On the next page are tables similar to Tables C-1 and C-2 with blanks substituted for the $L_i$'s. Please enter the loss values that you suggest into the tables on the next page.

Table C-3

Matrix of Loss Values Associated with Objectives of Primary Concern.

| Classification Decision | True Classification | |
|---|---|---|
| | Mastery | Non-Mastery |
| Mastery | ____ | ____ |
| Non-Mastery | ____ | ____ |
| Indeterminable | ____ | ____ |

Table C-4

Matrix of Loss Values Associated with Objectives of Secondary Concern

| Classification Decision | True Classification | |
|---|---|---|
| | Mastery | Non-Mastery |
| Mastery | ____ | ____ |
| Non-Mastery | ____ | ____ |
| Indeterminable | ____ | ____ |