A Model for Computerized Adaptive Testing Related to Instructional Situations

Stanley J. Kalisch
Educational Testing Service, Atlanta

The present study involved the formulation and evaluation by computer simulation of a model for computer-based adaptive testing related to instructional or training situations. Specifically, the model adresses tests composed of items corresponding to hierarchically related instructional objectives. The purpose of the endeavor was to formulate and to analyze a model that would reduce testing time without compromising the necessary level of accuracy in decisions regarding the mastery or nonmastery of objectives.

The adaptive testing model developed in this study combines the models of Ferguson (1969, 1970) and Kalisch (1974a, 1974b). Ferguson's procedure employs the Wald probability ratio test (Wald, 1947, 1973) to determine mastery/nonmastery of hierarchically related objectives. Kalisch's procedure employs a process that predicts item responses based upon prior examinees' data. For the present 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/nonmastery of objectives.

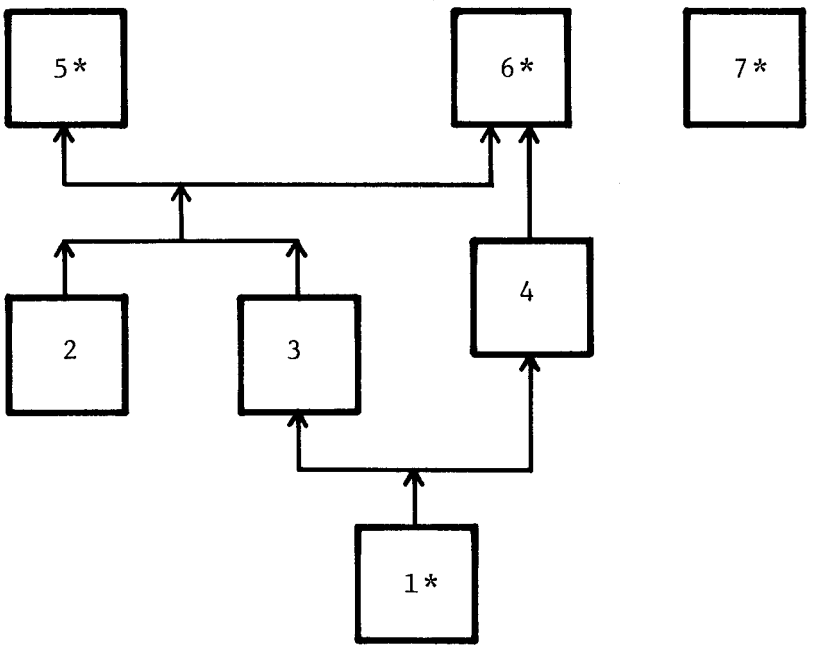
The Adaptive Testing Model

Configuration and Relative Importance of the Objectives

A hierarchical configuration of objectives, such as in Figure 1, defines the interrelationship of the objectives to be mastered by each trainee. 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. Nonmastery of either or both Objectives 2 and 3 implies nonmastery 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 direct assessment of performance on

Figure 1
Hypothetical Hierarchical Configuration of Objectives
(* Indicates an "Objective of Primary Concern.")



all the objectives was desired, then every objective would be identified as an objective of primary concern.

The model assumes that mastery of an objective implies mastery of all its immediate subordinate objectives; nonmastery of an objective implies neither mastery nor nonmastery 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. Nonmastery classification on an objective of primary concern results in testing each immediate subordinate as if it were also an objective of primary concern.

Basing Decisions on a Data Base

The decisions made in the adaptive testing process are dependent upon information collected from prior examinees. Although the existing model 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.

Two response-matching procedures were defined. With the first method a vector $\vec{s}$ of dichotomuously scored responses is generated for an examinee; for each additional response collected within a test, the $\vec{s}$ vector increases. The individual's $\vec{s}$ vector is matched with sets of responses in the data base; but only data base sets with exactly the same $\vec{s}$ vector (the same pattern of "1's" and "0's" 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/nonmastery 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.

Predicting item response correctness/incorrectness. Based upon the dichotomously scored responses to items presented to an examinee, conditional probabilities for answering the item correctly or incorrectly are determined on the basis of response patterns in the data base matching the examinee's. If either conditional probability exceeds prespecified levels, the correctness/incorrectness of the examinee's expected response is assumed.

Selection of items for presentation. Based upon an examinee's response pattern and the subset of the data base response matching the examinee's, items that are expected to provide the most information about the objectives of primary concern are selected for presentation. Two selection criteria were investigated in this study: item-objective agreement and inter-item agreement. For each method a coefficient was computed for each item not presented and for which prediction of correctness/incorrectness had not yet occurred. The item with the highest coefficient was presented to the examinee.

For the item-objective method, a coefficient of agreement between item i and the n objectives of primary concern was calculated as follows:

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

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

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

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;
$\bar{i} = 0$ means item $\bar{i}$ 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/nonmastery classifications for the examinee; and
$\vec{s}$ is the vector of the examinee's dichotomously scored item responses.

For the inter-item method a coefficient of agreement between item i and the n other items corresponding to the objectives of concern is computed according to the following formula:

$$A(i;i_1,i_2,\ldots,i_n) = [\{\sum_{j=1}^{n} [\text{Prob}(i_j = 1)|(\vec{r},\vec{s},i = 1)] \times [\text{Prob}(i = 1)|\vec{r},\vec{s}]\} + \{\sum_{j=1}^{n} [\text{Prob}(i_j = 0)|(\vec{r},\vec{s},i = 0)] \times [\text{Prob}(i = 0)|\vec{r},\vec{s}]\}/n \quad [2]$$

where

$i_j = 1$ is the probability of answering item i_j correctly;
$i_j = 0$ is the probability of answering item i_j correctly;
$\vec{r}$ is the objective mastery-nonmastery pattern for the examinee; and
$\vec{s}$ is the item response pattern (correct/incorrect) for the examinee.

Examinee response inconsistencies. "Untrue" responses by an examinee are those responses that do not agree with the examinee's "true" response (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

1. Selecting the correct answer by guessing, when in actuality the examinee should have answered the item incorrectly;
2. Providing an incorrect answer because of misinterpretation of part of the question; and
3. Pressing an unintended key on a terminal keyboard.

Item responses that are provided by an examinee, but 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. Vector $\vec{s}$ affects predictions of other item responses and selection of items for presentation. Generally, it is expected that item prediction errors will affect the accuracy of the system, whereas errors in item selection will reduce the efficiency of the system. Prediction and selection errors may 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 those made by the present exam-

inee, it would be expected that for each item the majority of prior examinees had 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 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 was selected, it was necessary that at least two items be presented for the examinee's responses prior to making predictions or to making 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. In this study the number of items presented was sufficient so that the probability of answering all of them correctly by chance alone was less than or equal to .5.

The purpose of obtaining responses to a set of independently selected items was to determine whether the examinee has demonstrated sufficient consistency in his/her 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 was computed as follows:

$$R_x = \frac{\sum_i G(x,i)}{\sum_i I(x,i)}, \qquad [3]$$

where

$$G(x,i) = \begin{cases} 1 \text{ if both responses to item } x \text{ and item } i \text{ were correct} \\ \quad \text{or if both responses were incorrect} \\ 0 \text{ if one response was correct and the other was wrong,} \end{cases}$$

and

$$I(x,i) = \{[\Sigma \text{ Prob}(i = 1 | x = 1)] \times \text{Prob}(x = 1)\} + \{\text{Prob}(i = 0 | x = 0) \times \text{Prob}(x = 0)\} . \qquad [4]$$

$G(x, i)$ was computed on the basis of the examinee's responses to item x and all the other items presented.

R_x indicates the examinee's consistency as compared to prior examinees' consistency. It is possible that a given examinee demonstrated greater consistency than prior examinees, but when the examinee's consistency was less than that for prior examinees, his/her item response pattern contained "untrue" responses. In this study the criterion for sufficiently consistent responses by an examinee required that for each item x, $R_x \geq .90$. If the criterion was not attained for each item, the item with the lowest R_x value was 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 the criterion). If $\vec{s}$ contained fewer elements, other items must be independently selected $\vec{s}$. Whenever the number of elements in $\vec{s}$ equaled or exceeded the minimum requirement, item selections and predictions were based upon $\vec{s}$. After the presentation of each additional item, all items for which responses were ob-

tained were included in the calculations of the R_x values. Hence, although an item response may be questioned and not included in $\underset{\rightarrow}{s}$, a future recalculation may indicate the item response to be consistent with the examinee's other responses. Likewise, items once contained in $\underset{\rightarrow}{s}$ may be excluded on a future recalculation.

Determining Mastery/Nonmastery of Objectives

For an objective of primary concern, the dichotomously scored results to all its items for which correctness/incorrectness has been determined or predicted were used with 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) were then used in the following formula:

$$S = \left(R \times \log_{10} \frac{C_f}{C_p} \right) + \left[(N - R) \times \left(\log_{10} \frac{1 - C_f}{1 - C_p} \right) \right] \qquad [5]$$

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 nonmastery score (difficulty of the objective for nonmasters);
C_p = the critical mastery score (difficulty of the objective for masters).

Mastery/nonmastery classifications were determined by comparing the value of S to ratios involving α and β (Type I and Type II errors); α is the error associated with falsely classifying an examinee as a nonmaster, and β is the error of falsely classifying an examinee as a master:

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

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

3. If neither of the above conditions was true, no mastery/nonmastery classification was possible (and additional item responses were necessary).

The model assumes that the classification of an objective for which insufficient items exist for a mastery/nonmastery decision is "indeterminate." This decision occurred whenever the pool of available items was exhausted before a mastery/nonmastery decision could 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 was approached by the examinee's proportion of items answered correctly. Ferguson (1969) used this procedure, but only after asking for 30 item responses for the objective. It appears that if an examinee cannot demonstrate mastery performance within a realistically expected number of items, immediately prescribing remedial instruction would 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 discriminations, or unrealistically high specifications for acceptable misclassification errors.

The adaptive testing procedure terminated when either of the following conditions occurred: (1) all objectives were classified as mastered or unmastered; or (2) the number of prior examinee observations in the data base upon which predictions are based was less than two. For the first condition, the test was terminated. For the second condition, unpresented and unpredicted items corresponding to objectives of concern were randomly presented to the examinee. Termination of the test occurred when each objective was classified.

Eight Versions of the Adaptive Testing Model

The adaptive testing model formulated for this study was applied in a 2 × 2 × 2 configuration of options. These derive from three options, each with two conditions: (1) two methods of item selection based upon item-objective agreement and inter-item agreement; (2) two response matching procedures based upon only item response patterns (only $\vec{s}$) and upon both item response and objective classification patterns (both $\vec{r}$ and $\vec{s}$); and (3) a dichotomous option regarding examinee response inconsistency. Table 1 provides a delineation of the options used for each version; the numbers used in the remainder of the report refer to combinations of options employed.

Phase I: Monte Carlo Simulations

The purpose of this phase of the study was twofold: (1) to test for the relative accuracy and efficiency of the eight versions of the adaptive testing model and a control version and (2) to study the relation of loss to individuals' achievement levels for the adaptive testing versions. Accuracy was examined in terms of correct mastery/nonmastery classifications. Efficiency was investigated in terms of the number of items presented to examinees.

The control version to which the adaptive testing versions were compared involved the testing of every objective. For each objective a prespecified number of items was randomly selected for each examinee. Under the control treatment, examinees generally received different items for an objective, but each received the same number of items. For each objective a randomly selected integer between 3 and 6, 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

Table 1
Options Employed in the Eight Versions
of the Adaptive Testing Model

Testing Version	Item Selection Method	Response Matching[1] Procedure	Inconsistency Check
1	Item-objective	Only $\underline{\vec{s}}$	No
2	Inter-item	Only $\underline{\vec{s}}$	No
3	Item-objective	Both $\underline{\vec{r}}$ and $\underline{\vec{s}}$	No
4	Inter-item	Both $\underline{\vec{r}}$ and $\underline{\vec{s}}$	No
5	Item-objective	Only $\underline{\vec{s}}$	Yes
6	Inter-item	Only $\underline{\vec{s}}$	Yes
7	Item-objective	Both $\underline{\vec{r}}$ and $\underline{\vec{s}}$	Yes
8	Inter-item	Both $\underline{\vec{r}}$ and $\underline{\vec{s}}$	Yes

[1] $\underline{\vec{s}}$ is the item response vector and $\underline{\vec{r}}$ is the objective mastery/nonmastery classification vector.

than N-1 resulted in a nonmastery 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 Air Force Base, Denver, Colorado.

Item response generation. 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.

For each examination by adaptive test version, two sets of examinee data were generated--one representing past examinees' responses and the other including responses that would be obtained from present examinees. For the control version, only one set of 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 testings. 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: (1) hierarchical configuration of the objectives; (2) objective parameters, such as difficulty; (3) discrimination, and passing criteria; (4) proportion and type of hierarchical errors; and (5) guessing factor for answering items correctly.

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

$$P\ (u = 1) = \begin{cases} d + \dfrac{\theta - \overline{\theta}}{1 - \overline{\theta}}\ (1 - d) & \text{if } \theta \geq \overline{\theta} \\ \\ d + \dfrac{\theta - \overline{\theta}}{\overline{\theta}}\ d & \text{if } \theta < \overline{\theta}\ , \end{cases} \qquad [6]$$

where

$P(\underline{u} = 1)$ = the probability of answering the item correctly;
$\underline{d}$ = difficulty of the item;
$\overline{\theta}$ = examinee's objective score; and
θ = mean objective score of the corresponding mastery/nonmastery group.

A random number $\underline{r}$ in the closed interval 0 to 1 was selected. If $\underline{r} \leq P(\underline{u} = 1)$, 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 examinee would attempt to guess the correct answer, given that his/her "true" response would be incorrect, was derived by the formula

$$P_1 = g_1(1 - \theta d) \qquad [7]$$

where

$\underline{g}_1$ is the risk factor for the examinee (from GENTAB);
θ is the examinee's objective score; and
$\underline{d}$ is the item difficulty for the examinee's mastery or nonmastery group.

A random number $\underline{r}$ in the interval 0 to 1 was selected. If $\underline{r}_1 \leq P_1$, the examinee would attempt to guess the correct answer. The probability of guessing correctly was obtained from the formula

$$P_2 = g_2 + g_2\ \theta d \qquad [8]$$

where g_2 is the guessing factor for the item (the probability of randomly selecting the correct answer), and θ and $\underline{d}$ are the same as defined previously. For all items, $\underline{g}_2$ was set equal to .2, assuming five alternatives to each item. A random number $\underline{r}_2$ in the interval 0 to 1 was selected. If $\underline{r}_2 \leq P_2$, the examinee was credited with answering the item correctly.

Experimental Design

The design employed 90 cells comprised of an element from each of the fol-

lowing two dimensions (independent variables): (1) Testing Version (8 adaptive test versions and 1 control test version) and (2) Examination (10 examinations). For each testing version, data were simulated for 50 hypothetical examinees, each of whom was to take 10 examinations using only 1 testing version across the 10 examinations. Hence, there were 450 hypothetical examinees, each taking 10 examinations.

Separate split-plot factorial analyses of variance were conducted for each of two dependent variables. The dependent variables were (1) total loss associated with errors in mastery/nonmastery classifications and (2) total number of items presented.

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 Table 2 represent the relative amounts of loss attributed to each mastery/nonmastery/indeterminate decision made, given the "true" mastery/nonmastery status.[1] It can be seen in Table 2 that under the known true situation of mastery, the best decision was to classify performance on an objective as "mastery." The positive numbers for decisions of "nonmastery" and "indeterminable" indicate there were errors involved with these decisions--the greater error being associated with the latter. Total loss equals the sum of the separate losses incurred for each objective decision for an examinee.

Table 2
Matrix of Loss Values Provided for
Objectives of Primary and Secondary Concern

Classification Decision	True Classification: Mastery	True Classification: Nonmastery
Objectives of Primary Concern		
Mastery	0	10
Nonmastery	5	0
Indeterminable	7	3
Objectives of Secondary Concern		
Mastery	0	6
Nonmastery	4	0
Indeterminable	5	2

Total number of items presented. Items for the adaptive tests were presented to provide information for predicting correctness/incorrectness of other

[1] Roger Pennell of the Air Force Human Resources Laboratory at Lowry Air Force Base provided losses based upon values independently obtained from individuals knowledgeable of the Air Force Weapons Mechanics training program.

items. The total number of items presented refers to the number of items answered by an examinee in order to make mastery/nonmastery decisions on objectives.

Experimental model. The split-plot factorial model used was

$$X_{ijkm} = \mu + A_i + B_j + \pi_{k(i)} + AB_{ij} + B\pi_{jk(i)} + \varepsilon_{m(ijk)} \quad [9]$$

where

X_{ijkm} is the dependent variable;
A_i is the testing version;
B_j is the examination; and
$\pi_{k(i)}$ is the subject effect.

A posteriori tests. With regard to the testing version 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 treatment. 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 sets of responses. For each of the 90 testing versions by examination cells, 50 hypothetical examinees were used.

α and β levels. In this phase of the study, the values of α and β relative to the Wald procedure were set at .2 and .1, respectively.

Results

All of the adaptive testing versions were significantly more efficient than the control version. Only one adaptive testing version demonstrated significantly smaller losses than the control version. An analysis of variance indicated significant Examination and Testing Version × Examination effects ($\alpha = .05$). A quasi-F statistic was computed for the testing version, 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 3 shows the results of the analysis of variance, and Table 4 provides the descriptive statistics for each testing version.

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 were reduced to one. The test effect remained significant at the .05 level, but the Treatment × Test interaction did not.

Dunnett's test indicated that the only adaptive testing version signifi-

Table 3
Analysis of Variance For Total Loss

Source	df	Mean Square	F
Between Subjects	449		
Testing version	8	623.32	1.14
Subjects-within-groups	441	525.15	
Estimates for quasi-F calculations	457	544.624	
Within Subjects	4050		
Examination	9	754.74	36.61*
Testing version × examination	72	40.09	1.94**
Examination × subjects-within-groups	3969	20.62	

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

cantly different (α = .05) from the control test was the sixth version--Adaptive Testing Version 6, using the inter-item agreement, based only on the item response vector, and employing the inconsistency check. Although the obtained t value of Adaptive Testing Version 7 did not exceed the critical value, the difference in the two was extremely small. The losses obtained for both versions were extremely close. Adaptive Testing Version 7 used item-objective agreement, based on both item response and objective classification vectors, and employed the inconsistency check.

Table 4
Descriptive Statistics of Total Loss for Each Examinee per Testing Version

Testing Version	Mean	SD	Range Min	Range Max
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

The Newman-Keuls test indicated no pattern of significantly different losses among the examinations. Although significant differences did occur between some pairs of examinations, no trend was indicated. The Testing Version × Exam-

ination 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 conventional test, except for the third examination.

For the number of items presented, an analysis of variance indicated significant Testing Version, Examination, and Testing Version × 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 5 shows the results of the analysis, and Table 6 provides the descriptive statistics for the number of items presented.

Table 5
Analysis of Variance For Number of Items Presented

Source	df	Mean Square	F
Between Subjects	449		
Testing version	8	31285.58	256.06*
Subjects-within-groups	441	2.56	
Estimates for quasi-F calculations	72	122.18	
Within Subjects	4050		
Examination	9	276.51	142.53*
Testing version × examination	72	121.56	62.66*
Examination × subjects-within-groups	3969	1.94	

*$p < .01$.

The results of the Newman-Keuls tests for the testing version effect showed that each adaptive test required significantly fewer ($\alpha = .05$) items than the control test. There were no significant differences among the adaptive versions.

Although significant differences existed in numbers of items presented for the 10 examinations, the adaptive testing versions varied only slightly in their relative efficiency. A version that appeared to require the fewest items on one examination may have required the most on another examination. The differences in the number of items required by the adaptive versions for any one test were not substantially different.

Loss as a Function of Achievement Levels

Although Adaptive Testing Version 6 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

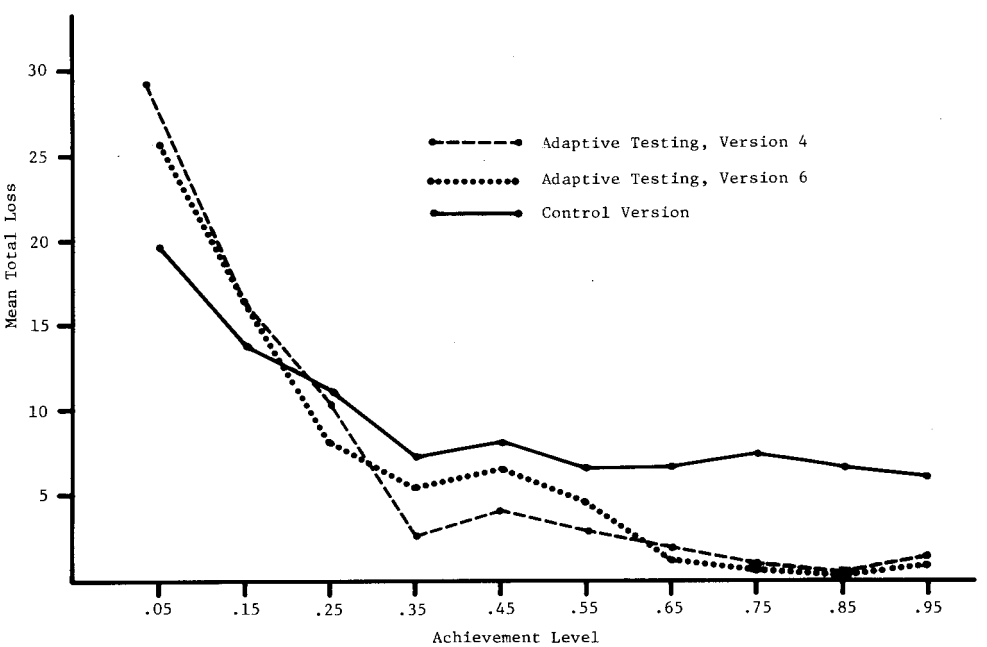
Table 6
Descriptive Statistics for Number of Items Presented

Testing Version	Mean	SD	Range Min	Range Max
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
Control	26.90	4.35	20	34

levels are shown for Adaptive Testing Versions 4 and 6 and for the control testing version in Figure 2.

The comparison of losses with respect to achievement levels demonstrated that both adaptive testing versions performed equally well throughout the

Figure 2
Mean Total Loss Corresponding to Levels of Achievement



achievement range. Version 6 demonstrated a slight advantage over Version 4 in the lower end of the achievement levels. This was probably due to the consistency check employed in Version 6.

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 α level. It may be recalled that β was set to .1, whereas α was set at .2. Since the false nonmastery error would be larger than the false mastery error, a higher proportion of false classifications would be expected for those 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 only a small proportion of examinees in the data base 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.

Selection of Adaptive Testing Versions for the Next Phase

The intention of the next phase of the study was to compare the results of some of the adaptive testing versions with those obtained in the present testing system used in the Air Force Weapons Mechanics training program at Lowry Air Force Base. Adaptive Testing Versions 4 and 6 were selected. No version was significantly superior in numbers of items presented. Adaptive Testing Version 6 was selected because of its superior accuracy. Adaptive Testing Version 4 was selected, however, solely on the basis of the mean number of items presented for item prediction.

Phase II: Real Data Simulations

Purpose

The purpose of this phase of the study was to compare (1) the relative efficiency of Adaptive Testing Versions 4 and 6 with each other and with the present testing method used in the Weapons Mechanics training program and (2) the classification decisions made from the adaptive testing versions with those made by the pesent method used in the Weapon Mechanics training program.

Design

The control testing version for 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 treatment.

Classification decisions made by the adaptive testing and control testing versions 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 was obtained when both indicate "nonmastery." Since for the adaptive tests, performance classified as "indeterminate" dictated procedures identical to those classified as "nonmastery," this condition was also considered an agreement. The α and β values selected were the same as in the previous phase--.2 and .1, respectively.

Data that were actually collected on four examinations in the Weapon Mechanics training program were used in the computer simulations for this phase.

For each 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 tests taken by the examinees. For each examination, the first 150 response sets, sorted in ascending chronological order, were used to form the data base. Of the remaining subjects, 50 were randomly selected as the examinees who were to take the simulated adaptive tests. Hence, within each examination the same 50 trainees were used as examinees, regardless of the testing version; but the same 50 trainees were not used across examinations.

The assumed hierarchical configurations for the objectives for each examination were provided by Roger Pennell of the Air Force Human Resources Laboratory, Lowry Air Force Base, Denver, Colorado. The mastery score for an objective with $N(\geq 2)$ items was set to $N - 1$, as is presently done with conventional testing procedure, which is referred to here as the control testing version. If N equaled 1, the cutting score was set to 1.

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 each adaptive testing version to the control testing version.

Results

Both adaptive testing versions used in this phase of the study demonstrated that each required significantly fewer items than the control testing version. Version 4 of the model required the presentation of fewer items than Version 6.

Efficiency. Adaptive Testing Version 4 required statistically significantly ($t = 8.30$, $df = 199$, $p < .001$) fewer items than Version 6. The descriptive statistics for these versions are shown in Table 7. Although there was a statistical difference, the superior efficiency of Version 4 amounted to less than one item per examinee per examination.

Table 7
Descriptive Statistics for Adaptive Testing Versions 4 and 6

Variable and Statistic	Adaptive Testing Version 4	Adaptive Testing Version 6
Number of Items Presented		
Mean	3.02	3.92
SD	1.19	1.42
Index of Agreement		
Mean	6.15	5.54
SD	3.39	3.27

Mastery/nonmastery decisions. Adaptive Testing Version 4 had a statistically significantly ($t = 5.58$, $df = 199$, $p < .001$) higher agreement in mas-

tery/nonmastery classifications than Version 6. The descriptive statistics for these versions are also shown in Table 7.

The average number of objectives per examination was 7.25. Hence, the range of the index could be from -7.25 to 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, Versions 4 and 6 had 92% and 88% agreement with the control testing version, respectively.

Separate t tests were performed on the number of items presented for each of the adaptive testing versions compared to the number required by that of the control version. The mean number of items presented under the control testing version across the four tests was 15.25. The number of items required by the adaptive testing version are presented in Table 8. The visual comparison of the tabled values reveals such large differences that no statistical test was necessary.

Since the four examinations differed in hierarchical configurations, number of objectives, and number of available items, Table 8 presents the percent of reduction in test items required by the adaptive testing versions in relation to the control testing version for each examination. The table also shows the percent of agreements in mastery/nonmastery decisions between each adaptive testing version and the control testing version.

Table 8
Comparison of Results of Adaptive Testing Versions 4 and 6 to Control Testing Version for Each Examination

Adaptive Testing Version and Examination	Number of Items Presented		Percent of Item Reduction	Number of Objectives	Percent of Mastery and Non-Mastery Agreements
	Control Version	Adaptive Testing Version			
Version 4					
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
Version 6					
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

The results show that both Adaptive Testing Versions 4 and 6 made most of the same mastery/nonmastery decisions as were presently being made by the Air Force in its Weapons Mechanics program; but the adaptive testing versions make the decisions with approximately 75% fewer items than the conventional, or control, version.

Discussion and Conclusions

Both simulation phases of the study have shown that the adaptive testing versions could make mastery/nonmastery decisions much more efficiently than testing on each objective with a constant number of items for each objective presented.

The real-data simulation showed that the mastery/nonmastery agreement between the control testing version and the adaptive testing versions was higher for Adaptive Testing Version 4. This does not mean that Version 4 is more accurate than Version 6. On the contrary, in the first simulation it was demonstrated that Adaptive Testing Version 6 was the only adaptive procedure that had significantly smaller loss than the control version. In essence, Adaptive Testing Version 4 and the control version in the second simulation phase would be expected to be equally as inaccurate in mastery/nonmastery decisions. Adaptive Testing Version 6 would be expected to be more accurate than the control version and, hence, would have fewer agreements with the control version than would Version 4.

Although in both phases statistically significant differences were found among the adaptive testing versions, the assignment of different values to the version's parameters might equalize all results. All the adaptive testing versions were used with the same values specified for the model's parameters. For example, for all versions, α and β were set at .2 and .1, respectively. The versions may be differentially sensitive to the parameters. Hence, two versions may be expected to perform exactly the same, but only by specifying different values for the same parameters.

For both simulation phases of the study the number of sets of responses needed in the data bases were unknown. For the second simulation phase it was estimated that 150 sets would be sufficient. The results indicate that an average of 29 sets matched each examinee's set on each test. The average of 29 sets per examinee did not give sufficient information as to whether the data base was of sufficient size. The ranges in number of sets indicated that for every test and for every adaptive testing procedure the data base was completely depleted for some examinees. As in the first simulation, it may not be that the data base contained insufficient numbers of response patterns but that there was an insufficient number of patterns for poorer 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 mastered the objectives, as in the Weapons Mechanics program, such a data base is insufficient for predictions of performance by nonmastering examinees. Hence, in such a situation, oversampling of nonmastering 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 adaptive testing versions in the monte carlo simulations and the superior efficiency demonstrated by Adaptive Testing Versions 4 and 6 procedures in the real-data simulations, it appears that any of the adaptive testing variations used in this study would be much more efficient than the conventional testing procedure used by the Air Force.

REFERENCES

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

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

Hays, W., & Winkler, R. Statistics, Volume 1: Probability, inference, and decision. New York: Holt, Rinehart, & Winston, 1970.

Kalisch, S. J. A tailored testing model employing the beta distribution 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 doctoral dissertation, Florida State University, 1974. (b)

Wald, A. Sequential analysis. New York: Dover Publications, 1973. (Originally published, 1947.)

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

ACKNOWLEDGMENTS

This study was sponsored by the Air Force Human Resources Laboratory, Air Force Systems Command, United States Air Force, Brooks AFB, San Antonio, TX 78235.