Adaptive Branching in a Multi-Content Achievement Test

Roger J. Pennell
D. A. Harris
Air Force Human Resources Laboratory
Technical Training Division

In an environment such as that offered by the Advanced Instructional System (AIS), the potential benefits derivable from adaptive testing become a practical reality. The AIS is an advanced development program to develop a computer-based educational and training system for the Air Force. The heart of the system is a CDC Cyber 70 which currently manages the training process for four courses at Lowry Technical Training Center through the so-called type "A" and "B" terminal. The type "A" terminal is an interactive plasma display terminal with graphic capabilities, while the type "B" terminal has test form reading and scoring capabilities, along with a line printer for issuing student prescriptions. The system is designed to manage the self-paced instructional process of a large number of students who spend approximately 40% of their time in a testing mode. Thus, with a large student flow through AIS courses requiring extensive testing, considerable benefits in terms of reduced training time are potentially available from procedures such as adaptive testing which reduce testing times.

Adaptive testing has also been called branched testing, response-contingent testing, sequential testing, and tailored testing. In the present study, the general term adaptive testing will be used to characterize any attempt to match test items to examinees based on a response history; the goal will be either reducing testing time or obtaining more valid and/or more reliable ability estimates.

Background

Realizing the potential of adaptive testing in a system such as the AIS, the Human Resources Laboratory initiated a multi-phase research study, beginning with the identification of a suitable algorithm to drive an adaptive testing program. The first phase identified the flexilevel approach of Lord (1971a, 1971b) as the tentative algorithm (Hansen, Johnson, Fagan, Tam, & Dick, 1974). Flexilevel testing has a number of advantages over other methods of adaptive testing. Namely, it is easily implemented, it does not require a large item pool, and it theoretically requires only $(n+1)/2$ items (where n is the number of items in the total test pool) to test each examinee. For example, a 25-item test would require only 13 items to test each examinee.

The flexilevel test first administers the item of median difficulty (difficulty levels are ascertained from pretesting). If an item is answered incorrectly, the next easiest unanswered item is given. If an item is answered correctly, the next most difficult unanswered item is given. An examinee continues testing until $(n+1)/2$ items have been answered.

Phase II of the above research effort was conducted in the Inventory Management (IM) course. The Block II test was used for the implementation of this study. The purpose was to validate the flexilevel adaptive testing paradigm with the primary goal of reducing test time. Each student was individually entered into the test, given the flexilevel adaptive test, and then given all remaining items.

A Phase III study was performed in Blocks II and IV of the Precision Measurement Equipment Course (PME). A task analysis was used to group items into five scales and to construct a hierarchy of scales within the test. The intent was to explore the feasibility of adaptive testing both within and across scales.

Study I

Objective

The purpose of the present study was to explore the kinds of conclusions which might be made by simulating flexilevel testing on paper-and-pencil protocols and comparing the results, i.e., estimated parameters, to those data actually collected on the computer terminal (Phase II). The intent was to evaluate the extent to which actual implementation and testing of the model on a computer terminal can be avoided.

A number of simulation studies of adaptive testing have been conducted, e.g., Cleary, Linn, & Rock, 1968a, 1968b; Paterson, 1962; Bryson, 1972; Linn, Rock, & Cleary, 1970. These studies have largely been concerned with ascertaining the potential benefits derivable from an adaptive testing paradigm, rather than extrapolating simulated results to actual adaptive data as this study did. Basically, the question posed by the present study was, "Must one actually conduct a study such as Phase II to ascertain the feasibility of adaptive testing?" Furthermore, "To what extent do simulated results parallel results under actual PLATO testing conditions?"

Method

A sample of 186 paper-and-pencil protocols was obtained from Inventory Management/Material Facilities (IM/MF) Block II. The test was composed of the same items used in the Phase II experiment. The sample was divided into two equal parts: a calibration (C) and a validation (V) sample. The C sample was used to estimate parameters necessary to implement the flexilevel testing algorithm. These parameters were then validated on the V sample in order to evaluate the stability of various dependent measures. The parameters estimated were (1) the item difficulties, implying the item ordering for flexilevel presentation and (2) the regression parameters for converting the flexilevel score into an estimated total score.

The flexilevel score could have been used to make the necessary pass/fail decisions required in a criterion-referenced testing situation such as that found in Air Force technical training. However, for two reasons it was desirable to translate back to the total score metric (percent correct). First, this is the metric traditionally used to assign scores. Second, the extent to which the flexilevel score reproduces the total score is a prime dependent measure in evaluating the feasibility of flexilevel testing.

The flexilevel score was derived as follows:

Let A index the set of items taken under flexilevel testing; and let d_i, $i \varepsilon A$, represent the difficulty of the ith item expressed as percent of the C sample answering correctly; furthermore, let

$$S_i = \begin{cases} 1 \text{ if item } i \text{ is answered correctly} \\ -1 \text{ if item } i \text{ is answered incorrectly.} \end{cases}$$

Then, the flexilevel score for the jth examinee, on the items completed, was defined as

$$F_j = \sum_i s_i d_i \qquad [1]$$

Stated more simply, F_j was the sum of the item difficulties answered correctly minus the sum of the difficulties answered incorrectly.

Since the total score, X_j, was available as the sum of correct responses divided by the number of items in the item pool (n=25), the usual regression equation

$$\hat{X}_j = a + bF_j \qquad [2]$$

was used to estimate the total score and the associated error $|\hat{X}_j - X_j|$.

It should be noted that the usual flexilevel rule of administering $(n+1)/2$ items to each examinee was departed from in both the Phase II study and the present study. That is, testing for a particular examinee was terminated if he/she were either to take a more difficult item but had already answered all of the difficult items or were to take an easier item but had already taken all of the easy items. This decision rule was used because as a function of entering examinees at varying locations on the item hierarchy, one of the dependent measures was the number of items required to terminate testing.

The dependent variable analyzed in addition to those mentioned above (viz., effect of item hierarchy, variable entry, and error in reproducing total score) was classification error. For a range of criterion levels, the error rate was examined, using $\hat{X}_j$ to classify students as failing or passing relative to their known classification based on X_j.

In addition to the C and V samples, a third sample (N=100) was obtained by randomly selecting test protocols of students who had gone through Phase II testing on the computer. This was possible since at the completion of each flexilevel session (using the same stopping rule described above) all items on the 25-item instrument which had not been administered were given. Thus, complete item protocols were available on this cross-validation (CV) group.

One intention of the Phase II study was to explore the utility of adaptively entering examinees into the item hierarchy. The entry point was calculated using three aptitude tests which the students took before they entered training. It was thought that adaptive entry might further reduce testing time beyond the reduction attributable to taking only $(n+1)/2$ items. Unfortunately, the CV sample was obtained when monitors were having difficulty obtaining the aptitude scores; therefore, the majority of the sample was entered at the $(n+1)/2$th item.

The comparison of the flexilevel results in the CV group, using the parameters estimated in the C group, explored whether or not a feasibility study such as Phase II needed to be conducted. Theoretically, the only difference between the CV and C groups was the use of a computer terminal to administer the test. This assumes item independence in the sense that taking items in a different order would not affect the test score.

Results and Discussion

The item difficulties, along with the correct responses, for the 25 items under study, are presented in Table 1. The mean item difficulty, an estimate of the mean test score, was .804. Typically, criterion-referenced test items tend to be quite easy; however, one of these items was exceptionally difficult (Item 6). Eliminating Item 6 raised the mean to about .84, which indicates that approximately 16% of the sample missed an item of average difficulty. The difficulties in Table 1 implied the ordering of the items for the simulated flexilevel testing; equal item difficulties implied an arbitrary ordering.

Next, the regression parameters for Equation 2 were estimated. Regression estimates for entering the item hierarchy at Item 3, 5, 7, 9, 11, 13, and 15 were calculated. These estimates are presented in Table 2 along with the correlation (validity) between X, the total score, and F, the flexilevel score (see Equation 1). The farther down on the item hierarchy (easier items) students were entered, the more items were required to terminate the flexilevel algorithm. This was vividly displayed by the trend of the regression weights. That is, increasing the entry point reduced the constant term, a, and increased the importance of the b term corresponding to the flexilevel score. The validities beginning at Entry Point 7 were quite good, indicating a high degree of accuracy in predicting total score. However, the cross-validated validities were of more interest.

Table 3 presents the V and CV group validities along with the C group for comparison. It should be noted that $\hat{X}_j$, the estimated total score, was computed using the weights developed in the C group. The validities for the

Table 1
Item Difficulties and Scoring Key, Group C

Item	Difficulty	Key
1	.968	2
2	.936	4
3	.819	2
4	.851	4
5	.809	5
6	.468	5
7	.670	2
8	.819	3
9	.819	1
10	.638	4
11	.915	3
12	.777	4
13	.777	5
14	.862	1
15	.894	1
16	.840	2
17	.840	3
18	.840	5
19	.723	4
20	.862	4
21	.691	4
22	.819	4
23	.755	2
24	.926	1
25	.777	4

Table 2
Regression Weights and Validities, Group C

Entry Point	a	b	Validity
3	.714	.388	.654
5	.656	.509	.773
7	.617	.560	.847
9	.578	.612	.926
11	.555	.631	.952
13	.524	.661	.972
15	.503	.671	.981

V group were strikingly high. In some cases they were higher than the C group, which indicated that the error in utilizing "non-optimal" regression weights and item difficulties was essentially non-existent. Some shrinkage was encountered in the CV group. However, this shrinkage all but disappeared after Entry Point 11. This indicated that parameters developed on paper-and-pencil protocols cross-validate to results obtained by use of computer terminals for high entry levels.

Table 3
Validities by Entry Point

Entry Point	Group C	Group V	Group CV
3	65	75	60
5	77	78	69
7	85	87	79
9	93	93	83
11	95	95	93
13	97	97	96
15	98	98	98

Note. Decimal points omitted.

Since the items used to construct the flexilevel score were also used (together with additional items) to compute the total score, the validities reported in Table 3 are inflated to some extent. The total score was computed by summing 1's and 0's corresponding to a correct or incorrect item, whereas the flexilevel score was computed by summing weighted item difficulties. Doubtless, the weighted item difficulties have a minimum built-in correlation with the 1-0 protocol.

Table 4
Percent Items Required to Terminate Testing

Entry Point	Group V	Group C	Group CV
3	20	20	19
5	30	30	30
7	41	40	41
9	52	50	52
11	62	60	62
13	70	69	72
15	78	77	80

Table 4 presents the average percent of items needed to terminate the flexilevel algorithm as a function of entry point. For example, when entering at Item 5, all three groups required an average of 30% of the total 25 items (7.5) to terminate the algorithm. The differences between the C sample and the V and CV samples presumably reflect an increase in number of test items required by using non-optimal difficulties, and thus a non-optimal item hierarchy for flexilevel branching. However, this effect was decidedly minimal.

Table 5 presents, in terms of number of items, the average and absolute error made in predicting total score. For example, when each group entered at the 11th item, the estimated total score ($\hat{X}_j$) differed by an average of .9 of an item from the known total score (X_j). Similar to Table 3, these data show comparable results across the three groups entering at Item 11 and above.

Table 5
Item Error in Predicting Total Score

Entry Point	Group V	Group C	Group CV
3	2.0	1.7	1.9
5	1.7	1.5	1.8
7	1.5	1.3	1.4
9	1.2	1.0	1.3
11	.9	.9	.9
13	.7	.6	.7
15	.5	.6	.5

Table 6 shows the average percentage of error of classification across various criterion levels. For a criterion of .70, for example, if $\hat{X}_j \geq .70$ and $X_j \geq .70$ or if $\hat{X}_j < .70$ and $X_j < .70$, the jth student was properly classified. However, if $\hat{X}_j \geq .70$ and $X_j < .70$ or if $\hat{X}_j < .70$ and $X_j \geq .70$, there would have been a classification error relative to the criterion of 70%. The percent of these errors averaged over criterion levels .40, .44, and .96 is the statistic presented in Table 6. When the three groups entered at Item 3, the cross-validated percentage of errors was about 11.5%, which doubtless would be unacceptably high to most course designers. On the other hand, errors of 6 or 7% might be acceptable when balanced against the decrease in overall training time.

Table 6
Percent Misclassified by Entry Point

Entry Point	Group V	Group C	Group CV
3	14	11	12
5	11	10	11
7	10	8	9
9	8	7	9
11	6	6	7
13	5	5	6
15	4	4	5

Conclusions

Making any decision regarding the implementation of adaptive testing involves a tradeoff between potential gains versus potential losses. It has been shown that fairly substantial decreases in the number of test items required are obtainable with very accurate estimation of total score. (An empirical question remaining is whether or not there is a decrease in testing time associated with the decrease in test items.) The tradeoff is relative to the decision categorizing an examinee incorrectly as passing or failing based on a flexilevel score. The above results indicate that this type of error ranges from about 5 to 12%. It should be noted, however, that the criterion used to gauge this error was the total score; this is far from an ideal criterion. What is needed, of course, is the "true score," i.e., the unknown indicator of whether or not a student has accomplished the behavioral objective, which is imperfectly measured by the total test score. Lacking such an indicator, the total score was used; however, there is no reason why the flexilevel test could not be making more valid decisions relative to the "true score." Indeed, this is one of the theoretical benefits attributable to adaptive testing.

The foregoing data have indicated that for reasonably high entry points, parameters estimated from paper-and-pencil test protocols cross-validate remarkably well to groups actually tested at a computer terminal using a flexilevel algorithm. This suggests that feasibility studies running actual subjects may not be called for. Rather, simulated results based on paper-and-pencil protocols may lead to a quick decision regarding whether or not adaptive testing should be implemented.

Study II

Objective

The objectives of Study II were (1) to summarize the data collected under the Phase III contract effort and (2) to offer some conclusions concerning the efficacy of flexilevel testing in an on-going training environment. The analysis was, of course, constrained by the manner in which the contractor implemented the study. However, the present analysis takes a different approach to the data and arrives at slightly different conclusions.

Method

A sample of 133 Precision Measuring Equipment (PME) students who were block tested on the PLATO terminal was obtained. Of those 133 protocols, 61 were Block II tests and 72 were Block IV tests. Both block tests contained 40 items; however, the subject matter covered by the tests was quite different.

A task analysis was performed in order to construct a hierarchical structure for each test. The task analysis grouped items into five relatively homogeneous scales according to item content. The scales were then placed in a hierarchical structure based on the relationships defined by the task analysis.

Table 7

Items Comprising Scales and Difficulties for the Block II Test (Calibration Sample *N*=105)

Scale 1		Scale 2		Scale 3		Scale 4		Scale 5	
Item	Difficulty	Item	Difficulty	Item	Difficulty	Item	Difficulty	Item	Difficulty
11	.97	24	.97	15	.98	26	.89	34	.95
10	.96	14	.96	29	.94	25	.88	31	.94
6	.96	1	.95	21	.94	39	.88	36	.93
6	.95	5	.90	16	.93	27	.81	37	.90
12	.94	3	.90	20	.92	40	.81	32	.85
7	.92	2	.75	17	.89	28	.70	38	.84
8	.86	23	.74	18	.87			35	.77
		13	.72	19	.85			33	.63
		4	.70	22	.84			30	.51
Mean	.94		.84		.91		.83		.81

Table 8

Items Comprising Scales and Difficulties for the Block IV Test (Calibration Sample *N*=113)

Scale 1		Scale 2		Scale 3		Scale 4		Scale 5	
Item	Difficulty	Item	Difficulty	Item	Difficulty	Item	Difficulty	Item	Difficulty
15	1.00	1	.96	29	1.00	31	.98	38	.96
16	1.00	10	.90	26	.99	39	.88	4	.95
18	1.00	11	.88	24	.98	37	.88	14	.85
8	.96	5	.88	23	.97	34	.87	13	.84
21	.96	22	.82	25	.94	32	.82	28	.81
2	.92	35	.62	27	.83	33	.70	17	.70
19	.86	7	.61	30	.72	36	.69		
12	.81					40	.57		
20	.82								
3	.67								
6	.58								
9	.58								
Mean	.93		.81		.92		.80		.85

All students entered the test at the median difficulty item of the first scale and were presented items based on the flexilevel algorithm described in Study I. After completing the flexilevel portion of each scale, the students were given the remainder of the items and then started at the median difficulty item in the next scale. This procedure was continued until all five scales were completed.

Results

The items comprising the scales, along with their difficulties, are presented in Table 7 and Table 8. As in Study I, the items were quite easy; the scale mean difficulties ranged from .81 to .94 in Block II and from .81 to .93 in Block IV. The average difficulty of a scale did not necessarily correspond to the position of the scale within the hierarchy. That is, the scales were not ranked in the hierarchy based on average difficulty, but rather by content.

The variables of interest were the proportion correct during the flexilevel portion of the test (S_j) and the flexilevel score (F_j), the latter being modified slightly from Study I. If R is defined as the set of items correct from the flexilevel test, w as the set of incorrect items, and P_i the difficulty of the ith item as obtained from Tables 7 and 8, then

$$F_j = \sum_{i \varepsilon R} (1-P_i) - \sum_{k \varepsilon w} P_k , \qquad [3]$$

where $i \varepsilon R$ and $k \varepsilon w$ define the flexilevel score for the jth student. Additional variables of interest were the percent of items saved, the amount of time saved relative to taking the full 40-item test, and the remainder score (the score achieved on those items not taken during the flexilevel portion).

Table 9
Summary Statistics for Dependent Measures

	Block II			Block IV		
Score	Mean	*SD*	*r* with Total Score	Mean	*SD*	*r* with Total Score
Total Score	.85	.39		.82	.39	
S_j (Proportion Correct)	.82	.40	.98	.79	.37	.98
F_j (Flexilevel Score)	.56	.19	.98	.47	.16	.98
% Items Saved	30.4	.89	.96	24.6	.83	.91
Remainder Score	.94	.35	.72	.93	.16	.66

Table 9 contains the means, standard deviations, and correlations with total score for S_j, F_j, percent of items saved, and remainder score for Blocks II and IV. Both S_j and F_j were almost perfectly related to the

total score, as evidenced by the correlation of .98. This indicated that after taking about 70% of the items in Block II and 75% of the items in Block IV, the prediction of a student's total score from S_j or F_j was almost perfect.

It was surprising that the relatively crude measure S_j performed as well as F_j, which was intended to be the more sensitive measure. F_j takes into account the difficulty of the item the student takes: Correctly answering an item (i) which is relatively easy results in a relatively small increase in score ($1-P_i$), and relatively large increases occur for correct answers to a relatively difficult item. Incorrectly answering a relatively easy item (i) results in a relatively large decrease in score (P_i), while relatively small decreases occur for incorrect answers to relatively difficult items. However, within the context of the present study, both measures performed equally well.

It can be seen from Table 9 that the mean remainder score was substantially higher than the corresponding total score. This was to be expected; with relatively easy items, students tended to emerge from each scale after taking the most difficult item. Therefore, the remaining items tended to be the easiest items with an associated higher score. Since the items were relatively uniform in difficulty, S_j or F_j should have been a good estimator of the remainder score. In fact, the associated correlations were approximately .55 across blocks.

Two questions remain to be answered. First, can testees accurately be classified into mastery or non-mastery states based on scores (i.e., S_j and F_j) calculated from the smaller item set? Second, was there any actual *time* savings associated with the item savings? The data relevant to the first question are reported in the next section.

Classification analysis. Regression equations for predicting total score ($\hat{X}_j$) from both S_j and F_j were computed (Equation 2). The predicted scores ($\hat{X}_j$) were then compared to the students' observed score (X_j), and the number of correct and incorrect classifications was calculated. For both blocks the course-established criterion of 70% was used to define the cutoff. However, using the total score as the measure of mastery or non-mastery was subject to the same criticism raised in Study I, namely, that the total score is an imperfect measure of mastery, the (latent) trait of interest. The Block II and IV regression equations and classification analyses are presented in Table 10. As can be seen, the prediction of total score pass-fail from either S_j or F_j in Block II was almost perfect; that is, the predicted score ($\hat{X}_j$) misclassified only 1.6% of the sample.

In Block IV F_j classified testees somewhat more accurately than S_j, i.e., 97.2% versus 94.4%. However, the errors in classification based on S_j were

Table 10
Regression Equations and Classification Analysis

Block II	Block IV
Regression Equations	
$\hat{X}_j = .08 + .94\ S_j$	$\hat{X}_j = .03 + 1.0\ S_j$
$\hat{X}_j = .49 + .65\ F_j$	$\hat{X}_j = .48 + .72\ F_j$

Hit-Miss Analysis Using S_j

Total Score (X_j)	Predicted ($\hat{X}_j$) Pass	Fail
Pass	52	1
Fail	0	8
% Correct = 98.4		

Total Score (X_j)	Predicted ($\hat{X}_j$) Pass	Fail
Pass	57	4
Fail	0	11
% Correct = 94.4		

Hit-Miss Analysis Using F_j

Total Score (X_j)	Predicted ($\hat{X}_j$) Pass	Fail
Pass	52	1
Fail	0	8
% Correct = 98.4		

Total Score (X_j)	Predicted ($\hat{X}_j$) Pass	Fail
Pass	60	1
Fail	1	10
% Correct = 97.2		

Table 11
Time (in Minutes) to Complete Scales

	Block II $N = 55$		Block IV $N = 65$	
	Flexilevel	Remainder	Flexilevel	Remainder
1	7.56	3.13	9.14	1.62
2	5.27	0.58	16.25	1.62
3	15.25	2.42	4.05	0.84
4	12.10	1.03	16.48	2.40
5	12.51	1.98	6.83	.93
Total Time on Test	1.03 hrs		1.00 hrs	
Flexilevel Time	.88 hrs		.88 hrs	
Proportion Time Saved	.15 hrs		.12 hrs	

Note. Sample sizes reduced due to occasional computer failure during testing.

conservative, since they classified students as failing the block test when they had actually passed.

Time analysis. The second question, concerning real time savings associated with item savings, was a most critical question. The study by Waters (1975) showed that time savings from adaptive testing procedures are generally minimal; in an operational training environment, a primary concern is whether or not training time and dollars can be saved by adaptive testing.

Data were collected on the amount of time taken by each student to complete the flexilevel portion of the test, as well as the amount of time taken to complete the remainder of the test. These times were collected for each scale in the block tests.

Table 11 presents the mean times for Blocks II and IV. The flexilevel test reduced testing time by only 15% and 12%, respectively. The procedure of starting each student at the median item of each scale required a minimum of 27 items before the flexilevel test was completed. Moreover, as pointed out earlier, those items which were not taken in the flexilevel portion tended to be the easier items and thus were answered relatively faster.

Conclusions

The results of the analyses suggest several conclusions about the efficacy of flexilevel testing in an on-going training environment. First, the proportion correct during the flexilevel test (S_j) is as effective in predicting total score as the ostensibly more sensitive flexilevel score (F_j). This fact was reflected in the correlation between S_j and total score, as well as in the accuracy of mastery or non-mastery classification. In addition, S_j has the advantage of being in the metric that is most familiar to both students and instructors.

It was also concluded that the modest time savings (12 to 15%) was due to the parameters used to implement flexilevel testing. That is, entering at the median item requires the administration of at least 27 items before exit from the test. In addition, items not taken during the flexilevel test tended to be easier; this was evidenced by the remainder score, which would tend to decrease the time a student needed to complete these items. However, it should be pointed out that even a 15% time saving applied to the large number of students in AIS courses will, in the long run, reflect a significant time savings.

Finally, the selection of the parameters for this study leads to speculation about potentially realizable savings resulting from alternate flexilevel strategies. The following study was designed to investigate that problem.

Study III

Objective

The results of Study II were obviously contingent on the parameters chosen to implement the study. For example, testees always began on the median item of a scale and took all scales. An alternative was to use the flexilevel algorithm at the scale level as well as at the item level (i.e., if a scale were passed, the next hardest scale was taken; if a scale were failed, the next easiest was taken, and so on). Study I has shown that the simulation of the flexilevel algorithm on paper-and-pencil test protocols closely approximated results obtained during testing on a computer terminal. Therefore, using Study II test protocols, the effects of adaptive movement across scales on the various dependent measures was simulated. In addition to implementing the flexilevel algorithm across scales, the simulation considered two other variables. First, the depth or item entry level within a scale was varied in a fashion similar to that used in Study I. Second, this depth notion was extended to the scale level by varying the starting scale between the most difficult and easiest.

Table 12
Items Comprising Scales and Difficulties

Block II									
Scale 1		Scale 2		Scale 3		Scale 4		Scale 5	
Item	Diff	Item	Diff	Item	Diff	Item	Diff	Item	Diff
15	.98	29	.94	5	.90	18	.87	35	.77
11	.97	21	.94	37	.90	8	.86	2	.75
24	.97	12	.94	3	.90	19	.85	23	.74
14	.96	31	.94	17	.89	32	.85	13	.72
9	.96	16	.93	26	.89	38	.84	28	.70
10	.96	36	.93	39	.88	22	.84	4	.70
6	.95	7	.92	25	.88	27	.81	33	.63
34	.95	20	.92			40	.81	30	.51
1	.95								
Mean	.96		.93		.89		.84		.69
Block IV									
15	1.00	1	.96	5	.88	27	.83	36	.69
16	1.00	8	.96	11	.88	20	.82	3	.67
18	1.00	21	.96	37	.88	22	.82	35	.62
29	1.00	38	.96	39	.88	32	.82	7	.61
26	.99	4	.95	34	.87	12	.81	6	.58
24	.98	25	.94	19	.86	28	.81	9	.58
31	.98	2	.92	14	.85	30	.72	40	.57
23	.97	10	.90	13	.84	17	.70		
						33	.70		
Mean	.99		.94		.87		.78		.62

Because of the overlap in item difficulties between the original scales, the items were reordered into scales based entirely on the difficulty indices obtained in the calibration sample. The scales were formed by ranking the items according to difficulty and then forming scales with non-overlapping item difficulties. The position of a scale in the hierarchy was determined by the average difficulty of the scale. Table 12 contains the new scales for the Block II and Block IV tests.

Method

The 133 test protocols obtained during Study II were used as the data in this study. The simulation consisted of varying the levels of three parameters and measuring the effects on the dependent measures. The three parameters manipulated were: (1) scale pass criterion (SPC); (2) scale start (SS); and (3) scale entry level (EL). These are defined below.

Entry level (EL) was used in the same way as in Study I. It defined the item number within each scale where the flexilevel algorithm was started. EL was varied between 1 and 5. If EL=1, the most difficult item was given first; and if EL=5, the fifth most difficult item was given first. EL also defined the minimum number of items that had to be taken before testing within a particular scale was completed. For example, with EL=1 at least one item had to be taken. If it were passed, testing was complete for that scale; if failed, at least one more was taken (the next easiest), and so on.

Scale start (SS) defined the scale within which testing was started, and, thus, took the values 1-5. If SS=5 (the most difficult scale) or SS=1 (the easiest scale), only one scale needed to be taken, i.e., if the most difficult were passed, or the easiest failed, testing was complete.

When the flexilevel strategy was implemented at the item level, the 1-0 item score was used to define the next item to be given, i.e., a "1" implied a more difficult item and a "0" an easier one. In a real sense, this was the criterion for movement between items. In a similar vein, a criterion for movement between scales was needed. This was complicated by variable entry (EL), since EL=1 implied possible scale scores of 1.0, .50, .33, whereas other values of EL implied other ranges of scale scores. Therefore, SPC was not operationalized completely satisfactorily in terms of number of items answered incorrectly. SPC thus was varied between 0 and 3, where a particular value defined the maximum number of items which could be incorrectly answered in order to pass the scale.

The assumption of item independence, which was important in Study I, was also relevant in this study. Namely, a subject taking a particular item in a different order would give the same response as he/she gave in the original order. To the extent that this assumption is true, the results presented below reflect potentially obtainable outcomes from a variety of flexilevel strategies.

Simulations. The computer simulation was used to generate the values of various dependent variables for all possible combinations of the three parameters for both Block II and Block IV. The dependent variables were (1) percent items saved; (2) the percent classified correctly by S_j; (3) the

percent classified correctly by F_j; and (4) the correlations with total score for S_j and F_j.

Results and Discussion

Table 13 presents the results of the simulation runs for Block II. Similar to Study I, EL strongly affected the dependent measures. Since EL implied the minimum number of items a student must take, the percent of items saved varied inversely with this parameter, i.e., maximum items saved with minimum EL. Also, as EL increased, the predictiveness of S and F was increased. This was also expected, since as EL increased, the item composite upon which S and F was based increased in size and thus reliability. As predictability increases, the percent of testees correctly classified would be expected to increase. In fact, it did increase.

Table 13
Simulation Results for Block II

Parameter	% Saved	Class (S_j)	Class (F_j)	Correlations $R_{S,T}$	Correlations $R_{F,T}$
SPC					
0	67	.933	.932	.829	.840
1	67	.942	.945	.833	.854
2	68	.946	.948	.834	.851
3	69	.919	.942	.830	.845
SS					
1	63	.933	.936	.851	.872
2	61	.949	.948	.877	.893
3	66	.948	.953	.859	.869
4	71	.937	.952	.819	.829
5	80	.908	.921	.753	.774
EL					
1	88	.884	.883	.674	.691
2	77	.925	.934	.817	.833
3	66	.954	.966	.861	.877
4	58	.961	.966	.896	.911
5	50	.949	.961	.911	.925

Note. Results for each parameter are averaged over the values of the other two variables.

The striking aspect of Table 13 is the very large savings in items obtainable with various flexilevel strategies: this is particularly dramatic for EL. At EL=1 only 12% of the items were required to correctly classify nearly 90% of the testees. At EL=2 only 23% of the original items were required to classify over 90% of the testees. This contrasts with the Study II strategy which saved 30% in Block II and 25% in Block IV, while correctly classifying 98% and 96% of the testees, respectively. It was apparent that for only a modest decrease in correct classifications, a large increase in test items saved could be realized. If the relationship between items saved

and time saved found in Study I were extrapolated to the present results, a 36% savings in test time could be realized at EL=2.

The relationship of the other parameters to the dependent measures was less clear. SS would be expected to introduce a bow-shaped effect on the dependent variables, since (similar to EL) SS implies the minimum number of scales which must be taken to complete testing. At least three scales are implied by SS=3; SS=2 or 4 implies at least two; and SS=1 or 5 implies at least one. This effect can be seen to some extent in the classification functions and validities which increased to SS=2 or 3 and then decreased. For SPC there was little to choose from in terms of an optimal value. The results for SPC were perhaps idiosyncratic to the generally easy nature of the test items, i.e., varying SPC had minimal implications for all but the least prepared student.

Table 14 presents the simulation results for the Block IV test. Again, EL had the strongest effect on each dependent variable. Indeed, the pattern for Block IV was much the same as the pattern reported for Block II. Results for these blocks suggested that generally optimum values for the parameters were SPC=2, SS=3, and EL=3.

Table 14
Simulation Results for Block IV

Parameter	% Saved	Class (S)	Class (F)	Correlations $R_{S,T}$	Correlations $R_{F,T}$
SPC					
0	66	.895	.911	.809	.82
1	66	.888	.919	.814	.843
2	69	.886	.915	.818	.847
3	69	.884	.900	.809	.83
SS					
1	63	.887	.908	.823	.858
2	60	.906	.926	.862	.883
3	63	.906	.926	.846	.861
4	69	.894	.917	.812	.829
5	79	.848	.878	.721	.749
EL					
1	88	.853	.862	.639	.656
2	77	.898	.910	.820	.842
3	65	.895	.927	.856	.882
4	56	.895	.925	.868	.897
5	49	.899	.931	.879	.904

Note. Results for each parameter are averaged over the values of the other two parameters.

Table 15 presents the values of the dependent variables for the Block II and IV simulations using the parameter values indicated above. These

results indicate that by using approximately 48% of the items there was 100% classification accuracy in Block II and about 93% in Block IV. The correlations of both S and F with the total score were also quite high. This suggested that total score could be predicted very accurately from either score (a fact observed in the classification data).

Table 15
Simulation Results: SPC=2, SS=3, EL=3

Block	% Saved	Class (S)	Class (F)	$R_{S,T}$	$R_{F,T}$
Block II	54	1.00	1.00	.94	.95
Block IV	51	.93	.94	.91	.93

Conclusions

Study III has shown that large savings in items and, potentially, test time can be realized through the implementation of alternate flexilevel strategies. The conservative strategy adopted in Study II resulted in only modest item and time savings. However, even these modest savings can result in significant dollar savings when amortized over thousands of technical training students in just one year. Study III has shown that significantly greater savings can be realized with more efficient procedures in the form of optimal values for SPC, SS, and EL.

Conclusions

The overall conclusion from the three studies is that flexilevel testing with variable entry offers an easily implemented testing procedure with potential for significant dollar savings at minimal risk (in the sense of misclassification). Studies I and III, the simulation studies, show the potential power of implementing alternate strategies and the great regularity of the data obtained.

The results from Study I indicate the viability of simulating flexilevel testing on paper-and-pencil protocols to determine optimal entry levels, as well as potential item savings. This type of simulation can be accomplished prior to actual implementation, or the results from Study III can be used directly to guide the selection of an optimal flexilevel strategy.

In any event, it would seem appropriate to implement further flexilevel testing in technical training where the availability of computer terminals permits. For example, since in the Advanced Instructional System students spend 30 to 40% of their time in testing activities, it can be seen that significant training time reductions are potentially obtainable.

References

Bryson, R. Shortening tests: Effects of method used, length, and internal consistency on correlation with total score. Proceedings of the 80th Annual Convention of the American Psychological Association, Washington, DC: The American Psychological Association, 1972, 7-8.

Cleary, T. A., Linn, R. L., & Rock, D. A. An exploratory study of programmed tests. Educational and Psychological Measurement, 1968, 28, 345-360. (a)

Cleary, T. A., Linn, R. L., & Rock, D. A. Reproduction of total test score through the use of sequential programmed tests. Journal of Educational Measurement, 1968, 5, 183-187. (b)

Hansen, D. N. Johnson, B. F., Fagan, R. L., Tam, P., & Dick, W. Computer-based adaptive testing models for the Air Force technical training environment Phase I: Development of a computerized measurement system for Air Force technical training (AFHRL-TR-74-48), Brooks Air Force Base, TX: Air Force Human Resources Laboratory, July 1974.

Linn, R. L., Rock, D. A., & Cleary, T. A. Sequential testing for dichotomous decisions. (College Entrance Examination Board Research and Development Report No. 3, RDR-69-80). Princeton, NJ: Educational Testing Service, 1970. (ETS RB 70-31)

Lord, F. M. The self-scoring flexilevel test. Journal of Educational Measurement, 1971, 8, 147-151. (a)

Lord, F. M. A theoretical study of the measurement effectiveness of flexilevel tests. Educational and Psychological Measurement, 1971, 31, 808-813. (b)

Paterson, J. J. An evaluation of the sequential method of psychological testing. Unpublished doctoral dissertation, Michigan State University, 1962.

Waters, B. K. An empirical investigation of the stratified adaptive computerized testing model. Applied Psychological Measurement, 1977, 1, 141-152.