Multi-Content Adaptive Measurement of Achievement

David J. Weiss
and
Joel M. Brown
University of Minnesota

All the previous research in adaptive testing has been concerned with tests which covered only a single content area. Thus, all of the branching procedures implemented for adaptive selection of items to be administered to a testee have been designed exclusively for intra-test branching within a single, presumably unidimensional, content area. Unidimensional approaches to intra-test adaptive testing are useful for measurement in the achievement domain (e.g., Bejar, Weiss, & Gialluca, 1977; Bejar, Weiss, & Kingsbury, 1977). Frequently, however, achievement tests span several content areas. Consequently, in many cases the assumption of a single dimension may not be appropriate. For these kinds of achievement tests, or for achievement test batteries covering a number of separable content areas for which separate scores are required, none of the existing adaptive strategies (Weiss, 1974) are directly applicable.

There are two reasons why many of the adaptive testing strategies developed for single-content area ability tests may not be appropriate for achievement tests which cover several content areas. The first reason is that although the unidimensional branching models can be applied to separate content areas, they are not designed to take into account the information available between content areas. The second, and more practical, reason is that it might not be possible to generate relatively large numbers of items such as those required for many adaptive testing strategies within one content area in an achievement test. Urry (1977) has suggested that item pools to be used in adaptive testing with Owen's (1975) Bayesian testing strategy should include a minimum of 100 items to measure one dimension. Although there are no firm guidelines for other adaptive testing strategies, it is evident that they will function best with large item pools. Thus, application of these strategies to an achievement test battery of five subtests would require the test constructor to assemble 500 items with good psychometric qualities. Frequently, this is not possible. Consequently, in the application of adaptive testing to the unique problems in the measurement of achievement, an important research issue is the identification of adaptive testing strategies which make efficient use of existing item pools, rather than requiring the re-design of test item pools to meet the requirements of specific adaptive testing strategies.

The present paper describes an adaptive testing strategy which can be used in achievement tests with relatively small numbers of items. The strategy is designed for achievement test batteries or achievement tests with multiple content areas. It incorporates both intra-subtest branching and inter-subtest branching in order to efficiently adapt the test battery to each individual testee. The adaptive testing strategy is applied to a test battery and evaluated in terms of

1. The reduction in number of items administered,
2. Correlations of ability estimates with those derived from conventional administration of the test battery, and
3. The effects of adaptive administration on the psychometric information in the test scores.

METHOD

Purpose

The purpose of this study was to develop and evaluate an efficient and generalizable adaptive testing strategy for an achievement test battery comprised of a number of subtests. The adaptive testing strategy developed is designed to operate within a fixed item pool containing a relatively small number of items for each subtest. Real data simulation techniques (Weiss & Betz, 1973, pp. 11-12) were used. That is, the adaptive testing strategy was applied to item response data obtained from the administration of an achievement test battery which previously had been administered conventionally by paper-and-pencil. Results for the conventional testing strategy were compared with those for the adaptive testing strategy in terms of both test information and test length.

Procedure

Test Items and Subjects

Achievement test data were provided by the Personnel and Training Evaluation Program (PTEP) of the Naval Guided Missile School at Dam Neck, Virginia. These data were from a systems achievement test (SAT F17603) battery administered to 365 fire control technicians. The test battery included 12 subtests, each covering knowledge areas for different equipment or subject matter. Table 1 shows the content and number of items in each subtest. The test battery was administered in one booklet containing 232 items. The number of items per subtest ranged from 10 to 32; all of the items were multiple-choice with four response choices. The data provided by PTEP consisted of an identification number for each testee, the testee's number correct score on each of the 12 subtests, and correct-incorrect item responses for each of the 232 items.

Item Parameterization

Items were parameterized using Urry's ESTEM computer program for latent trait item parameterization employing the three-parameter normal ogive model (see Urry, 1976, p. 99). This program provided estimates of the item discrimination (*a*), item difficulty (*b*), and guessing (*c*) parameters. The items for each subtest were parameterized independently of items in other subtests.

Table 1
Number of Items in Each Subtest

Subtest	Content	No. of Items
A	Fire control system casualty procedures	10
B	Optical alignment group	10
C	Control console and power subsystem	18
D	Platform positioning equipment	22
E	Multiplexed equipment	18
F	Digital control computer and software	18
G	Digital control computer--operator interface	14
H	Magnetic disk file	12
I	Digital control computer--missile interface	24
J	Guidance and guidance testing	29
K	MTRE MKG MOD3	32
L	Spare guidance temperature monitor	25
Total		232

Adaptive Testing Strategy

The adaptive testing procedure was developed in order to reduce to a minimum the number of items administered to each individual with as little impact as possible upon the measurement characteristics of the test battery. Both intra-subtest adaptive branching and inter-subtest adaptive branching were used in the development of the procedure.

Intra-Subtest Branching

Item Selection. The basic concept for intra-subtest adaptive branching was that the order in which the items were to be administered was to be dependent upon values of the item information curve (Birnbaum, 1968, p. 462). For each item in each subtest, item information values were computed for values of θ ranging from -3.0 to +3.0 in steps of .2. Items were selected within a subtest for each testee by computing the value of all item information curves at the current estimated achievement level ($\hat{\theta}$) for that testee. The item selected for administration was the item which had the highest information value at the testee's current level of $\hat{\theta}$. Once an item was administered to a testee, it was eliminated from the subtest pool of available items for that testee.

Estimation of θ. Owen's (1975) Bayesian scoring procedure was used for this simulation study. This scoring procedure provides an achievement level estimate ($\hat{\theta}$) after each *m*th test item is administered. The procedure begins with a prior estimate of $\hat{\theta}_m$ and its variance (σ^2_m). For the first item of the first subtest administered (m=1), these were 0.00 and 1.00, respectively. An item was administered and scored as correct or incorrect. The revised estimate of $\hat{\theta}$ was determined using equations provided by Owen (1975, p. 353). The updated estimates of $\hat{\theta}$, along with their associated variances, were used as the prior estimates of $\hat{\theta}$ for the selection of the next test item, which was based on the maximum information rule described above. The next item was administered, and a new value of $\hat{\theta}$ was determined which was then used to select the next item. This procedure was repeated until a termination criterion was reached.

Termination criteria. Two criteria were used in determining when administration of items within a subtest should be stopped: (1) when all of the remaining items provided less than a pre-determined small amount of information or (2) when the within-subtest item pool was exhausted. Testing was terminated for a given testee at the first occurrence of one of these criteria within a given subtest. In applying the first criterion, testing was terminated when there was no item available which provided an information value greater than .01 at a given testee's current level of $\hat{\theta}$. Figure 1 diagramatically summarizes the intra-subtest branching procedure.

Figure 1
Intra-Subtest Branching Scheme

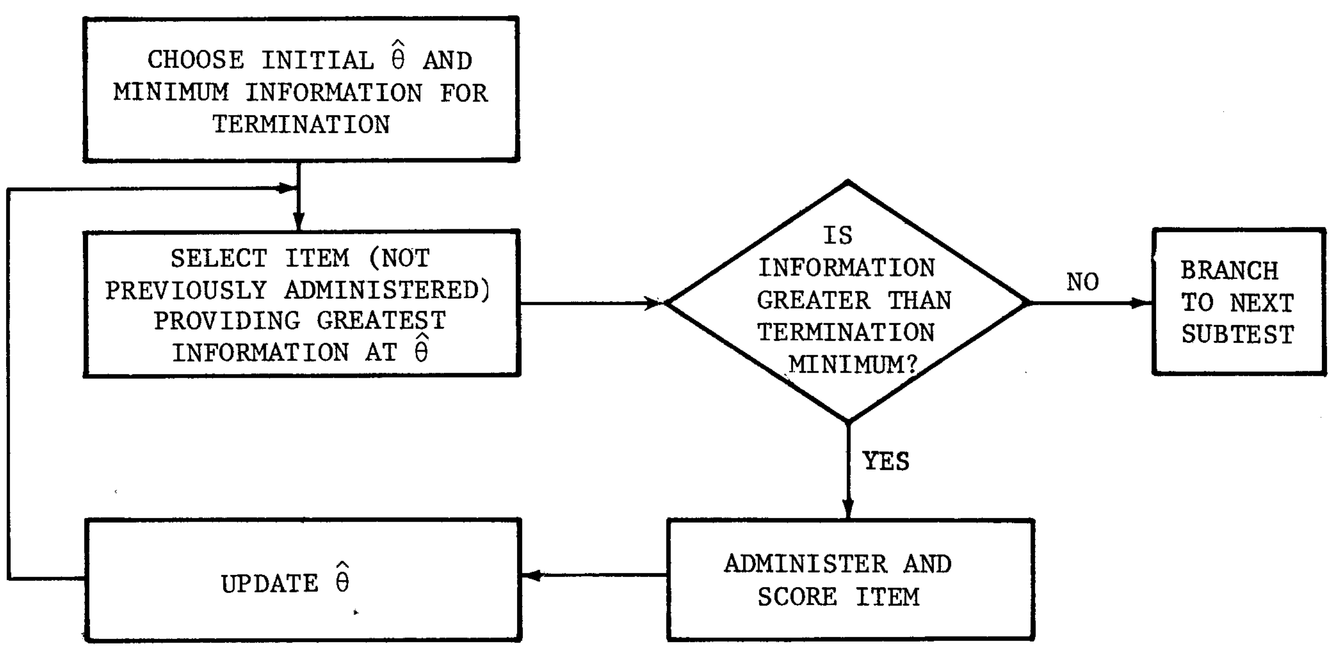
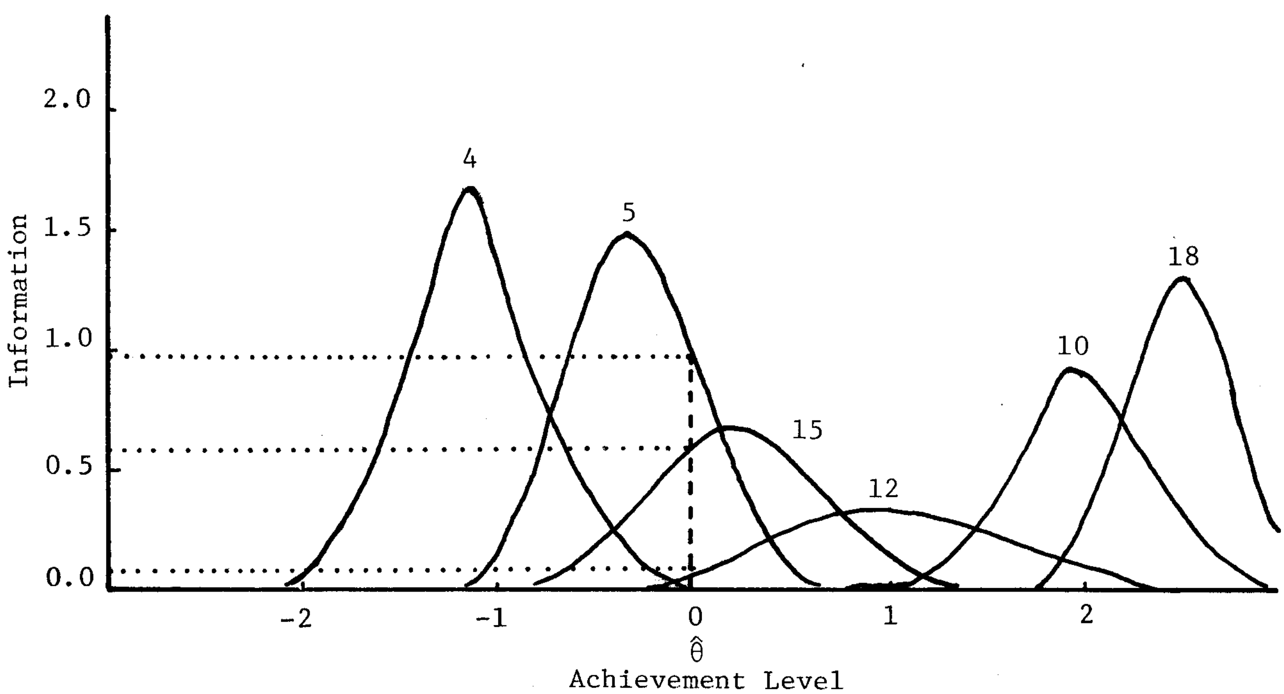


Illustration of Intra-Subtest Adaptive Branching

Illustrating this procedure, Figure 2 shows estimated item information curves for 6 items from Subtest 1. (There were a total of 15 items in Subtest 1 from which only 6 were chosen to simplify the illustration.) The height of the information curve at a given achievement level indicates the amount of information provided by the item. Most of the items are fairly "peaked"; that is, they provide information over a relatively narrow range of the achievement continuum. While the information curves overlap to some degree, different items provide different amounts of information at a given point on the achievement continuum. The guiding principle for the adaptive procedure was to administer the item which provided the most information at the current achievement estimate.

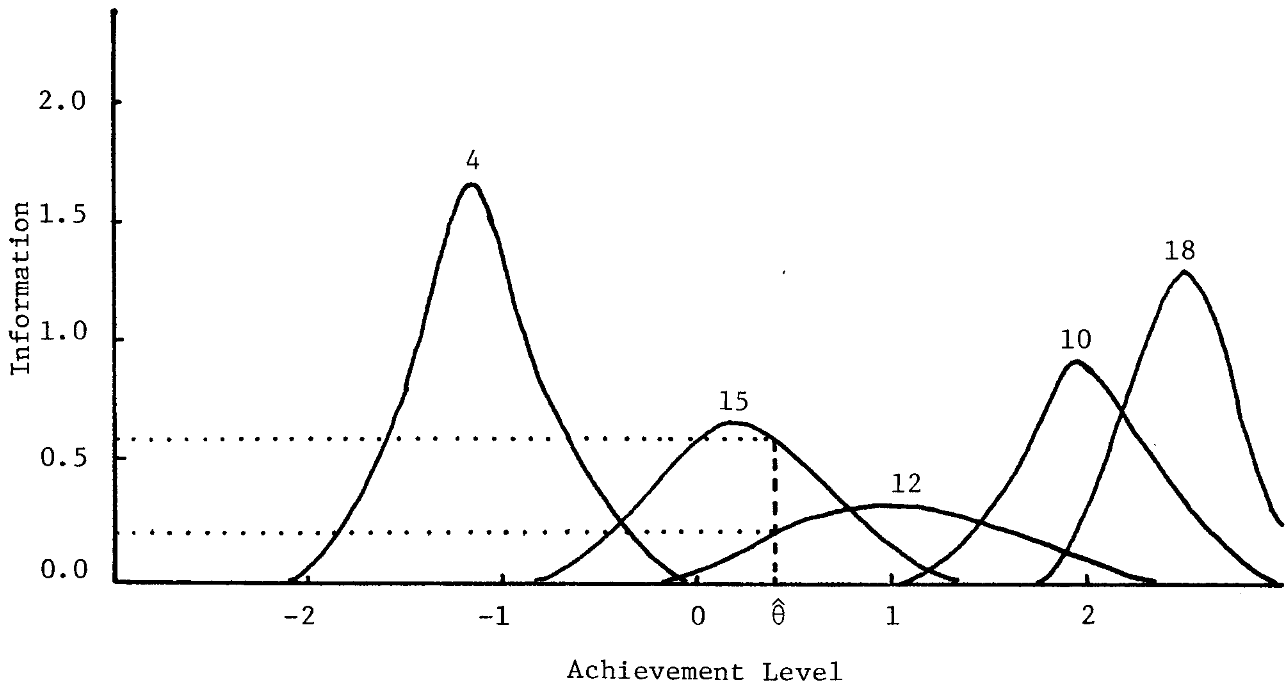
Figure 2
Estimated Item Information Curves for Six Items from Subtest 1



For a testee beginning Subtest 1, the initial achievement estimate was $\hat{\theta}=0$ (for subsequent subtests, this varied by individual); this is shown by the vertical dashed line in Figure 2. Of the six items in the example, only three items had essentially non-zero information values at $\hat{\theta}=0$. These values, shown by the horizontal dotted lines in Figure 2 were .90 for Item 5, .58 for Item 15, and .04 for Item 12. Applying the rule that the item selected is the one which provides the most information at the current $\hat{\theta}$, Item 5 would be selected for administration.

Figure 3 shows the revised value of $\hat{\theta}=.46$ derived from the Bayesian scoring routine, assuming that a correct answer was given to Item 5. The information curve for Item 5, which was already administered, is not shown in Figure 3. At the new value of $\hat{\theta}$, only Items 15 and 12 provide non-zero values of information. Since Item 15 has an information value of .54 and Item 12 has a value of .20, Item 15 is selected as the second item to be administered to this testee.

Figure 3
Estimated Item Information Curves for Five Items from Subtest 1



Assuming that the testee had correctly answered Item 15, the value of $\hat{\theta}$ increased to .92; this is shown in Figure 4. At that value of $\hat{\theta}$, Item 12 provides .32 information and Item 10 provides .02 information. Item 12 is thus administered next. Assuming that Item 12 was answered incorrectly, the $\hat{\theta}$ decreased to .62, which is plotted in Figure 5. The figure shows that of the three items remaining, none provides any information at the current level of $\hat{\theta}$. Thus, there is no need for administering additional items from Subtest 1; and testing in that subtest is terminated. The achievement level estimate of $\hat{\theta}_1$=.62 is taken as the testee's score on Subtest 1, since it is based on all items providing more than non-trivial amounts of information about that testee's achievement level.

Inter-Subtest Branching

Subtest ordering. The order of administration for the various subtests was chosen to take maximum advantage of the intercorrelations among them, thereby utilizing the redundant information in previously administered subtests. This was accomplished through linear multiple regression. First, the number correct subtest scores for the 12 subtests were intercorrelated; and the

Figure 4
Estimated Item Information Curves for Four Items from Test 1

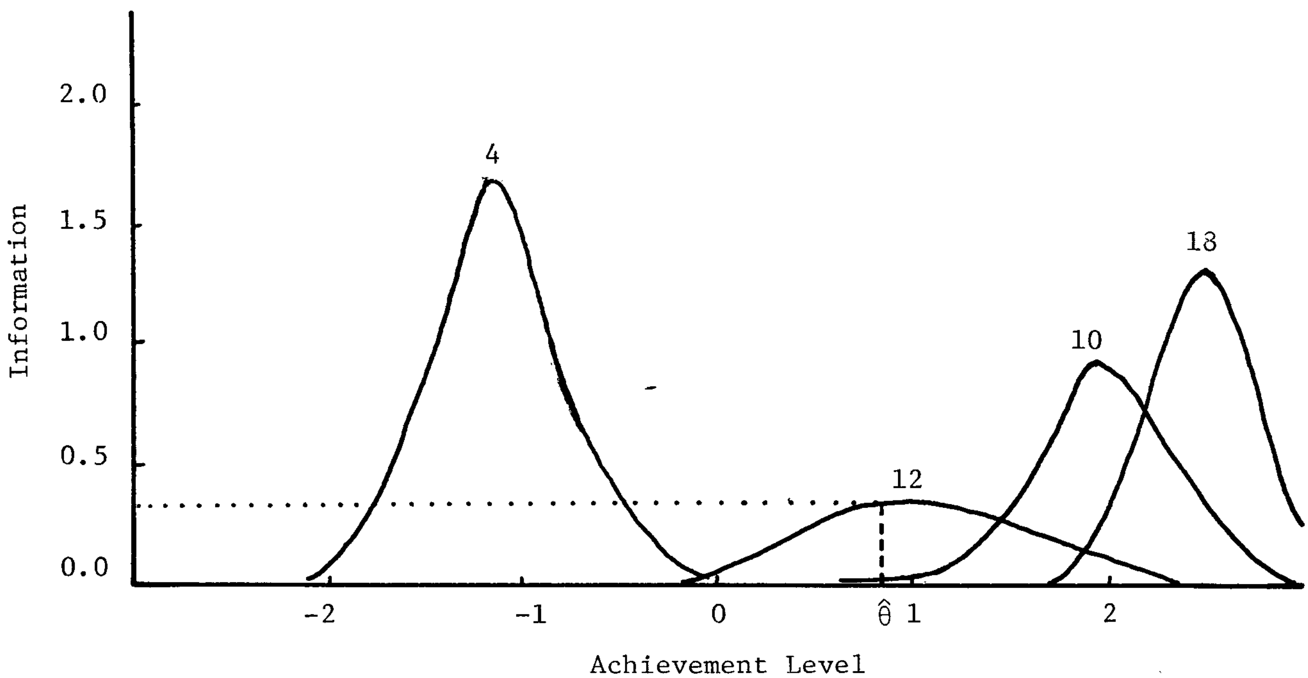
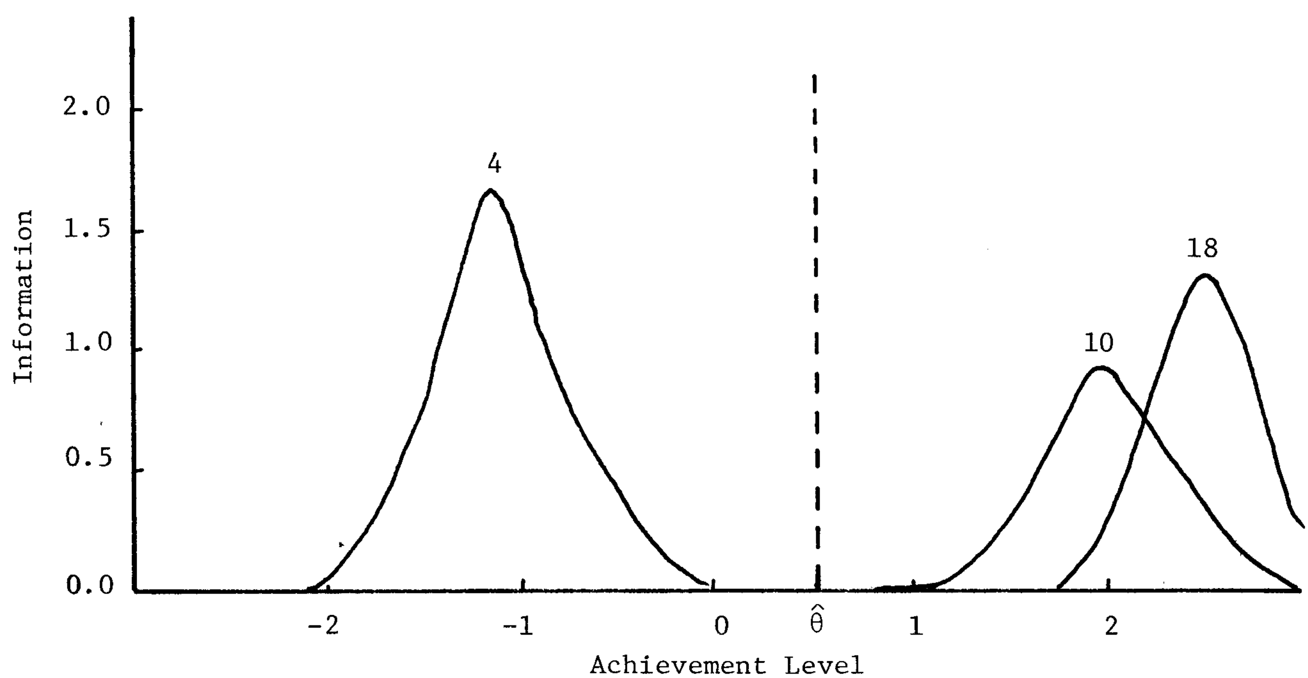


Figure 5
Estimated Item Information Curves for Three Items from Test 1



highest bivariate correlation was chosen from the intercorrelation matrix. One of these two subtests was arbitrarily designated to be administered first; the other was designated to be administered second.

Multiple correlations were then computed using the subtests previously designated first and second as predictor variables. Each of the 10 remaining subtests, in turn, was designated as the criterion variable. Of these 10 subtests, the one which had the highest multiple correlation with the first and second subtests was designated as the third subtest. This procedure was repeated to select the fourth subtest for the adaptive administration, computing multiple correlations with the first 3 subtests as predictor variables and each of the remaining 9 subtests, in turn, as the criterion variable. That subtest having the highest multiple correlation with the first 3 subtests was selected as the fourth subtest to be administered. By adding one subtest to the predictor set at each subsequent stage, this procedure was continued until all 12 subtests were ordered.

As a result of this procedure, the order in which the subtests were administered was the same for all testees. However, the selection of items within each subtest and the order in which those items were administered varied with testees as a function of the amount of item information provided at the testee's current achievement estimate.

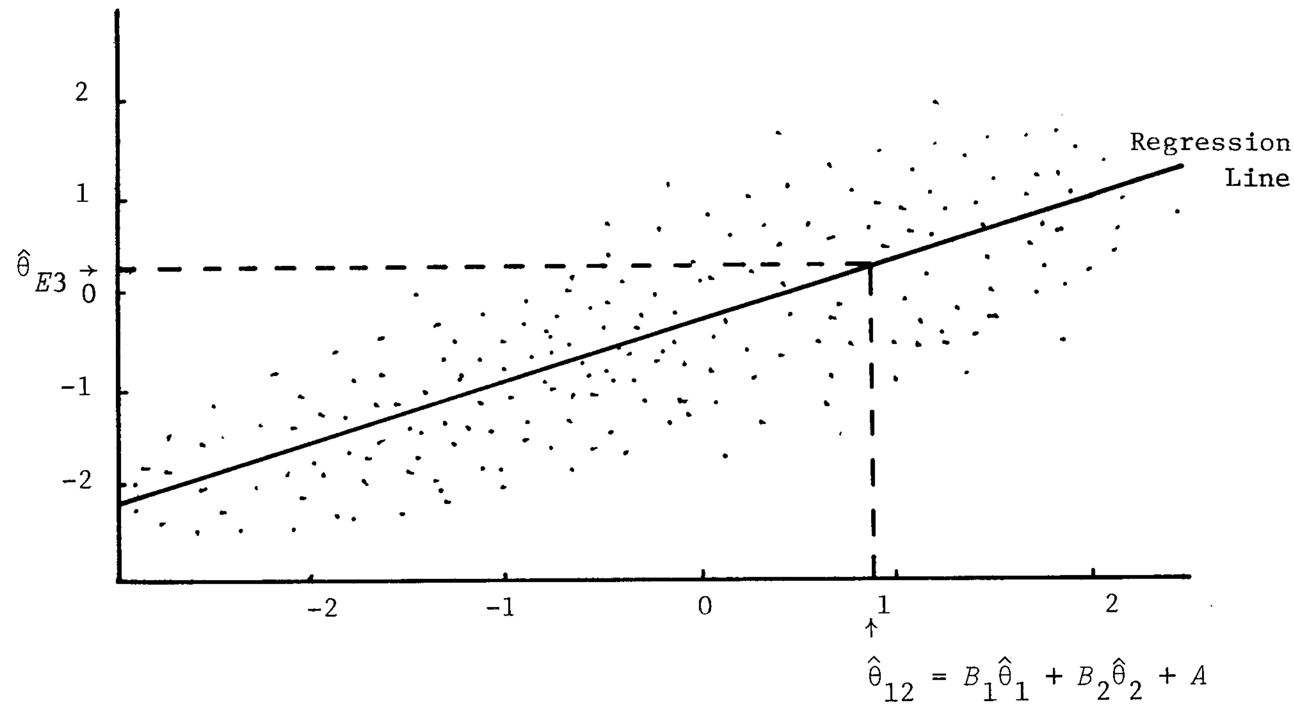
Differential subtest entry points. An important feature of the adaptive testing strategy implemented in this study was that after the first subtest, each testee's entry points for the second and subsequent subtests were differentially determined. For the first subtest, each testee's achievement level was assumed to be $\theta=0.00$. That is, having no previous information on which to base an estimate of the testee's achievement level, the initial item chosen from the first subtest for administration was the item which provided the most information for an estimated achievement level at the mean of the $\hat{\theta}$ distribution. Thus, all testees began the first subtest with the same test item.

The entry point into the item pool for the second subtest was determined from both the examinee's $\hat{\theta}$ at the end of the first subtest and the bivariate regression of scores from Subtest 1 on Subtest 2. This regression equation was based not only on scores for the items administered adaptively, but also on the correlations derived from number correct scores for all items in each of the subtests.

Determination of the entry point for the third and subsequent subtests was merely a generalization of the method used for the second subtest. The testee's achievement level estimates from Subtest 1 ($\hat{\theta}_1$) and Subtest 2 ($\hat{\theta}_2$) were entered into the multiple regression equation for predicting Subtest 3 scores from scores on Subtests 1 and 2. This generated an estimated subtest score for an individual ($\hat{\theta}_{E3}$) which was used as the initial prior achievement level estimate for intra-subtest branching in Subtest 3. The squared standard error of estimate from the multiple regression of Subtests 1 and 2 on Subtest 3 was used as the initial prior variance of the Bayesian achievement level estimate for Subtest 3. Figure 6 illustrates this differential entry point procedure.

Figure 6

Estimation of Initial Achievement Level Estimate for Subtest 3 ($\hat{\theta}_{E3}$)
From the Multiple Regression of Subtest 1 ($\hat{\theta}_1$) and Subtest 2 ($\hat{\theta}_2$)



The inter-subtest branching regression procedure was used for entry into each of the remaining subtests. Each subsequent regression equation was based on the achievement estimates from each of the previously administered subtests. A testee's achievement level estimates for each subtest, based on the multiple regression of all previous subtests on a new subtest, was used as the initial Bayesian prior $\hat{\theta}$ for intra-subtest branching within that subtest. Item selection and scoring within subsequent subtests was then based on the intra-subtest branching procedures described earlier.

Conventional Test

A conventional test was used for comparison with the adaptive testing strategy. The subtests were administered in the same order for both the conventional and adaptive strategies. In the conventional strategy, all items within each subtest were administered sequentially so that all testees took the same items in the same order. Hence, there was no differential entry for the conventional strategy. In addition, all testees completed all items, which is typical in conventional testing. In order to facilitate comparison of results with the adaptive strategy, Bayesian scoring was employed for the conventional test. A mean of 0.0 and a variance of 1.0 were used as the initial prior achievement estimate of the Bayesian score for each subtest.

Data Analysis

The basic question examined in this study was whether the number of items administered could be reduced through adaptive testing without significantly changing the characteristics of the test scores. The effects of reducing the number of items by the adaptive testing item selection procedure were evaluated by means of both a correlational analysis and an information analysis.

Correlation Analysis

Early research comparing single test adaptive testing strategies with conventional testing strategies (see Betz & Weiss, 1973, 1974; Larkin & Weiss, 1974, 1975; Vale & Weiss, 1975; Weiss, 1973) demonstrated that adaptive tests resulted in test scores highly correlated with conventional test scores, even though the adaptive tests required substantially fewer items. Consequently, in the present study Pearson product-moment correlations were computed between subtest achievement level estimates ($\hat{\theta}$) from the conventional and adaptive testing procedures in order to examine the extent of the relationship between the scores. These were computed separately for each of the 12 subtests. High correlations between the scores would suggest that the tests ranked the examinees in a similar order along the achievement continuum.

Information Analysis

Information analyses were conducted in order to compare the adaptive and conventional testing strategies as a function of achievement levels. Test information values for different testing strategies at different levels on the achievement continuum provide an indication of their relative degree of precision of measurement (Birnbaum, 1968). Estimated test information curves were generated separately for each subtest for both conventional and adaptive testing strategies.

In the conventional testing strategy, an examinee's subtest information value was computed by summing the item information values at the examinee's final estimated achievement level ($\hat{\theta}$) for that subtest. An estimated information curve was plotted for the total group of examinees from their individual achievement level estimates and corresponding information values. For a conventional test, this is equivalent to computing the test information function using the item parameters, a, b, and c, as suggested by Birnbaum (1968, pp. 454-464).

Estimated subtest information curves were generated similarly for the adaptive testing strategy. The estimated value of test information was computed at each testee's final achievement estimate for the subtest by summing the information values at that $\hat{\theta}$ for the particular subset of items administered to that testee. Thus, for both adaptive and conventional testing, each test information value was computed at the final value of $\hat{\theta}$ for the subtest, based on the information provided by the items actually administered.

RESULTS

Test Length

The number of items administered under both the adaptive and conventional test strategies is summarized in Table 1. The data in Table 1 show substantial reductions in test length as a result of the adaptive testing strategy. For Subtest 1, 15 items were administered by the conventional procedure,while from 4 to 13 items were administered by the adaptive procedure. Fifty percent of the group answered between 7 and 10 items in the adaptive test. The mean number of items administered by the adaptive strategy in Subtest 1 was 8.73, which represents a 41.8% reduction from the number of items required by the conventional test.

Table 1
Number of Items Administered in 12 Adaptive and Conventional Subtests

Subtest	Conventional Test	Adaptive Test				Percent Reduction[a]
		Mean	*S.D.*	Range Min	Range Max	
1	15	8.73	1.86	4	13	41.8
2	24	14.12	2.90	4	20	41.2
3	17	9.87	3.38	2	17	41.9
4	22	12.57	4.60	2	22	42.9
5	19	11.55	3.58	1	18	39.2
6	13	4.70	2.10	1	12	63.8
7	18	7.44	3.21	1	15	58.7
8	10	7.07	1.71	1	10	29.3
9	10	6.44	1.72	1	9	35.6
10	23	8.42	5.54	1	22	63.4
11	12	5.52	2.97	1	12	54.0
12	18	5.41	3.20	1	15	69.9
Mean	16.75	8.49	3.06	1.67	15.42	49.3
Test Battery	201	101.84	24.08	27	153	49.3

[a] Computed by the formula 100-[(Mean number of items in adaptive test/mean number of items in conventional test)×100]

Similar results were observed for the other subtests. Reduction of number of items required by the adaptive test varied from a low of 29.3% for Subtest 8 to a high of 69.9% for Subtest 12, in which a mean of 5.41 items was administered by the adaptive strategy. In Subtest 12, between 3 and 7 items were administered to 50% of the testees in the adaptive strategy as compared to 18 items for each testee in the conventional test. Subtest 12 had the highest percent reduction. In all probability, this was attributable to the increased accuracy of the test entry point from the multiple regression of the scores on the 11 prior subtests.

It is interesting to note that for Subtests 5 through 12, the minimum number of items administered by the adaptive procedure was one. For several of these subtests, a relatively substantial number of testees were administered only one item, i.e., almost 10% of the total group for Subtests 6, 11, and 12. The minimum number of items administered by the adaptive strategy was less for tests later in the adaptive testing sequence. This probably resulted from the increased use of prior test information for determining the initial item to be administered.

Although minimum numbers of items were administered at relatively high frequencies by the adaptive strategy, the maximum numbers of items were administered to very few testees. For Subtests 3, 4, 8, and 11 the maximum number of items administered by the adaptive strategy was the same as that administered by the conventional test; frequencies associated with these maximums were 2, 1, 5, and 1, respectively. For the remaining 8 subtests, none of the testees received the same number of items in the adaptive tests as they did in the conventional tests.

The conventional test battery consisted of 201 items administered to all testees. The average number of items administered by the adaptive strategy (see Table 1) was 101.84, representing a 49.3% reduction in number of items administered. The median number of items administered was 103, indicating a slight negative skew to the distribution. Fifty percent of the testees received between 86 and 119 items in the adaptive battery, representing reductions of 57.2% to 40.8% for half of the testees. None of the testees required all the items in the adaptive administration. The longest adaptive battery administered required 153 items for one testee, representing a 23.9% reduction in test length; the shortest adaptive battery for one testee required only 27 items, representing a test length reduction of 86.6%.

Correlation Analysis

Table 2 shows the Pearson product-moment correlations of the Bayesian achievement level estimates ($\hat{\theta}$) for the conventional and adaptive testing strategies: 11 of the 12 correlations were greater than .90. The highest correlations were .98 for Subtests 2 and 8; the lowest was .74 for Subtest 6.

Table 2
Correlation (r) of Bayesian Achievement Level Estimates ($\hat{\theta}$) for the Adaptive and Conventional Testing Strategies by Subtest, and Cronbach's Alpha Coefficient for the Conventional Subtests

Subtest	No. Items	r	Cronbach's Alpha
1	15	.91	.57
2	24	.98	.69
3	17	.96	.54
4	22	.97	.65
5	19	.93	.59
6	13	.74	.44
7	18	.90	.50
8	10	.98	.56
9	10	.95	.39
10	23	.92	.61
11	12	.91	.51
12	18	.94	.40

The items contributing to the Bayesian subtest achievement level estimates in the adaptive test were a subset of those used in the conventional test. Thus, to some extent, the magnitudes of the correlations in Table 2 were a function of this part-whole relationship. This is supported by a comparison with the Alpha internal consistency estimates for the conventional subtests shown in Table 2. If there were no part-whole relationship, the correlations between the achievement level estimates would be restricted by the internal consistencies. However, all the correlations were substantially higher than the Alpha values.

If the magnitude of the correlations of the two achievement estimates were primarily determined by the part-whole relationship attributable to common items, the number of items administered in a subtest would bear a strong relationship to these correlations. This was not generally the case: One of the two highest correlations (r=.98) was observed for Subtest 8, which had only 10 items in the conventional test, while Subtest 9, which also had 10 items, had an r=.95. Although Subtest 8 had the smallest percentage reduction attributable to the adaptive administration (20.3%; see Table 1), Subtest 9 had a 45.6% reduction and Subtest 2 (r=.98) had a 41.7% reduction. Subtest 6, which had the lowest r (.74), had a 63.8% reduction attributable to adaptive testing; but the highest percent reduction (69.9%) was observed for Subtest 12, for which an r=.94 was observed between the adaptive and conventional achievement estimates. Thus, these data suggest that the magnitudes of the correlations shown in Table 2 were not a direct function of either the number of items in the conventional tests or the internal consistency of those tests.

Information Analysis

Tables 3 and 4 provide mean raw values of estimated information $[I(\hat{\theta})]$ at intervals of $\hat{\theta}$ for the adaptive and conventional tests for ordered Subtests 1 and 12. These values are based on mean information in test items actually administered to each testee, using the testee's $\hat{\theta}$ at the termination of each subtest.

Figure 7 shows a plot of estimated information values from adaptive and conventional administration of Subtest 1; estimated information values for the last subtest administered, Subtest 12, are shown in Figure 8. The information obtained from the adaptive administration of Subtest 1, for all practical purposes, was identical to the information from the conventional administration. The largest mean difference in information between adaptive and conventional administration, .14, occurred in the estimated achievement interval between $\hat{\theta}$=-1.39 and $\hat{\theta}$=-1.20. For Subtest 12 the information curves resulting from adaptive and conventional administration were practically identical except that the adaptive strategy produced a wider range of estimated achievement levels. That is, for Subtest 12 there were 46 testees who obtained adaptive scores which were less than the lowest of conventional scores.

CONCLUSIONS

This paper has presented an adaptive testing strategy designed for use with achievement test batteries covering multiple content areas. One goal of

Table 3
Adaptive and Conventional Test Mean Information Values [$I(\hat{\theta})$]
and Mean Difference in Information and t Values
at Estimated Achievement Levels ($\hat{\theta}$) for Subtest 1

$\hat{\theta}$ Interval		Adaptive Test			Conventional Test			Mean Difference		
Min	Max	N	$I_a(\hat{\theta})$	$S.D.$	N	$I_c(\hat{\theta})$	$S.D.$	$[I_c(\hat{\theta})-I_a(\hat{\theta})]$	t	df
-3.00	-2.80	0			0					
-2.79	-2.60	0			0					
-2.59	-2.40	0			0					
-2.39	-2.20	0			0					
-2.19	-2.00	0			0					
-1.99	-1.80	0			0					
-1.79	-1.60	11	.70	.29	14	.64	.23	-.06	-.58	23
-1.59	-1.40	22	1.85	.40	23	1.83	.35	-.02	-.18	43
-1.39	-1.20	21	2.87	.04	25	2.73	.18	-.14	-3.49**	44
-1.19	-1.00	23	2.86	.04	20	2.89	.03	.03	2.75**	41
-0.99	-0.80	25	2.86	.06	28	2.89	.06	.03	1.82	51
-0.79	-0.60	33	3.36	.24	37	3.38	.19	.02	.39	68
-0.59	-0.40	21	4.15	.15	19	4.15	.16	.00	.00	38
-0.39	-0.20	31	4.21	.11	24	4.26	.06	.05	2.01	53
-0.19	0.00	27	3.72	.19	32	3.75	.23	.03	.54	57
0.01	0.20	35	3.02	.21	30	3.04	.21	.02	.38	63
0.21	0.40	26	2.43	.09	31	2.50	.12	.07	2.45*	55
0.41	0.60	42	2.17	.04	29	2.23	.08	.06	4.17**	69
0.61	0.80	14	1.90	.00	27	1.96	.07	.06	3.19**	39
0.81	1.00	10	1.85	.00	10	1.81	.04	-.04	-3.16**	18
1.01	1.20	13	1.74	.00	7	1.74	.01	.00		
1.21	1.40	0			6	1.85	.01			
1.41	1.60	11	2.10	.00	3	2.13	.00	-.06		
1.61	1.80				0					
1.81	2.00				0					
2.01	2.20				0					
2.21	2.40				0					
2.41	2.60				0					
2.61	2.80				0					
2.81	3.00				0					

*$p \leq .05$
**$p \leq .01$

the strategy was to select and administer items within a subtest as a function of the amount of information provided by each item at each testee's current estimated achievement level. A second goal was to use redundant information between and among subtests (by predicting a testee's performance on subsequent subtests based on performance on previous subtests) to determine appropriate differential entry points in adaptive branching between subtests. It was hypothesized that attaining these goals in the design of an adaptive testing strategy would result in considerable reduction in the number of items

Table 4
Adaptive and Conventional Test Mean Information Values [$I(\hat{\theta})$]
and Mean Difference in Information and t Values
at Estimated Achievement Levels ($\hat{\theta}$) for Subtest 12

$\hat{\theta}$ Interval		Adaptive Test			Conventional Test			Mean Difference		
Min	Max	N	$I_a(\hat{\theta})$	$S.D.$	N	$I_c(\hat{\theta})$	$S.D.$	$[I_c(\hat{\theta})-I_a(\hat{\theta})]$	t	df
-3.00	-2.80	0			0					
-2.79	-2.60	0			0					
-2.59	-2.40	11	.11	.32	0					
-2.39	-2.20	7	.04	.01	0					
-2.19	-2.00	15	.06	.03	0					
-1.99	-1.80	13	.20	.04	0					
-1.79	-1.60	12	.41	.07	1	.53				
-1.59	-1.40	15	.88	.28	10	.95	.17	.07	.71	23
-1.39	-1.20	23	1.73	.24	21	1.81	.26	.08	1.06	42
-1.19	-1.00	23	2.63	.67	24	2.81	.32	.18	1.18	45
-0.99	-0.80	17	4.04	.24	31	3.86	.28	-.18	-2.24*	46
-0.79	-0.60	27	4.57	.14	32	4.56	.12	-.01	-.30	57
-0.59	-0.40	33	4.64	.83	44	4.80	.01	.16	1.28	75
-0.39	-0.20	23	4.75	.02	35	4.76	.02	.01	1.86	56
-0.19	0.00	49	2.07	2.37	80	2.03	2.37	-.04	-.09	127
0.01	0.20	19	4.97	.09	24	4.36	1.69	-.61	-1.57	41
0.21	0.40	16	5.23	.05	16	5.23	.06	.00	.00	30
0.41	0.60	10	5.25	.05	12	5.27	.04	.02	1.04	20
0.61	0.80	11	4.84	.17	10	4.89	.15	.05	.71	19
0.81	1.00	10	3.35	1.77	9	4.29	.16	.94		
1.01	1.20	4	3.73	.04	7	3.60	.09	-.13		
1.21	1.40	7	2.89	1.28	5	3.36	.05	.47		
1.41	1.60	4	3.30	.00	1	3.32		.02		
1.61	1.80	1	3.32		0					
1.81	2.00	1	3.69		0					
2.01	2.20	0			0					
2.21	2.40	0			1	6.66				
2.41	2.60	1	6.67		0					
2.61	2.80	0			0					
2.81	3.00	0			1	2.26				

* $p<.05$

administered to each testee, while sacrificing little, if any, test information compared to that obtainable by administering the entire test battery conventionally. Thus, the focus of this adaptive testing strategy was utilization of an existing item pool for an achievement test battery to efficiently measure or estimate each testee's achievement level.

The adaptive testing strategy described in this report (see Brown & Weiss, 1977, for further analyses of these data) provides methods for intra-subtest and inter-subtest branching which exclude the administration of unnecessary items. The data indicate that on this achievement test battery the length of the battery can be reduced by 50% for the typical testee. In no case was it

Figure 7
Observed Information Curves for Subtest 1 (Conventional Test = 15 Items; Adaptive Test = 8.74 Items Average, Range = 4 to 13 Items)

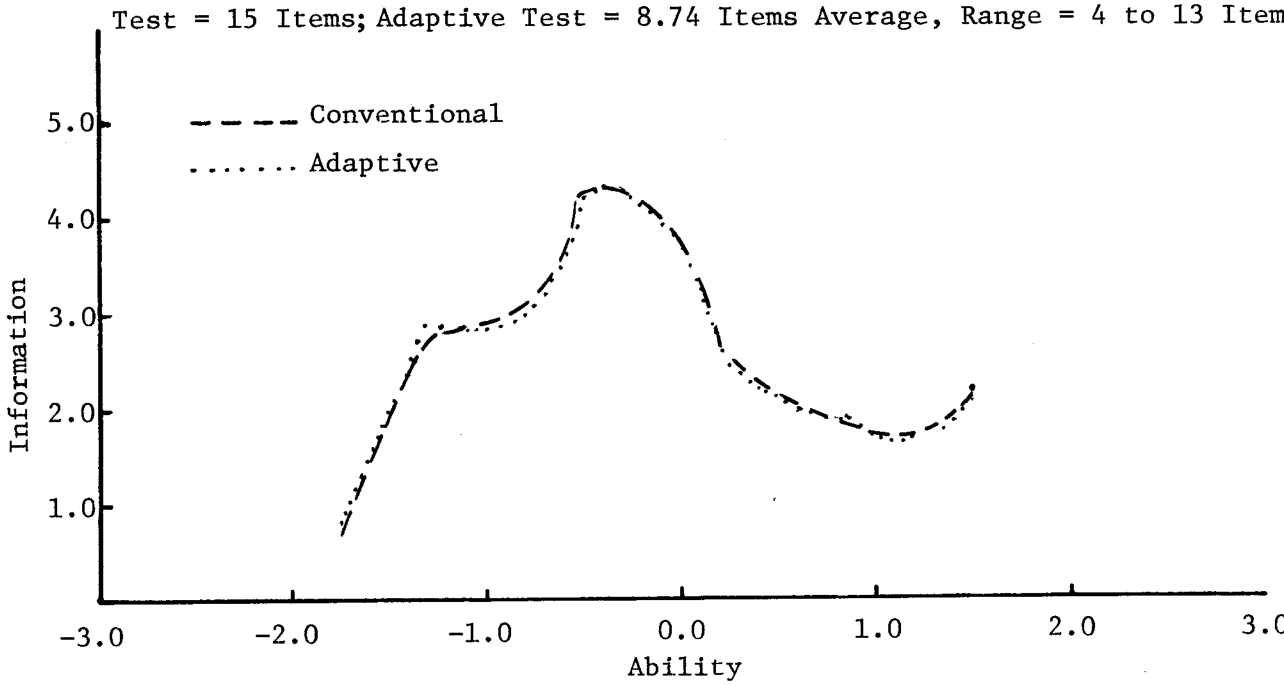
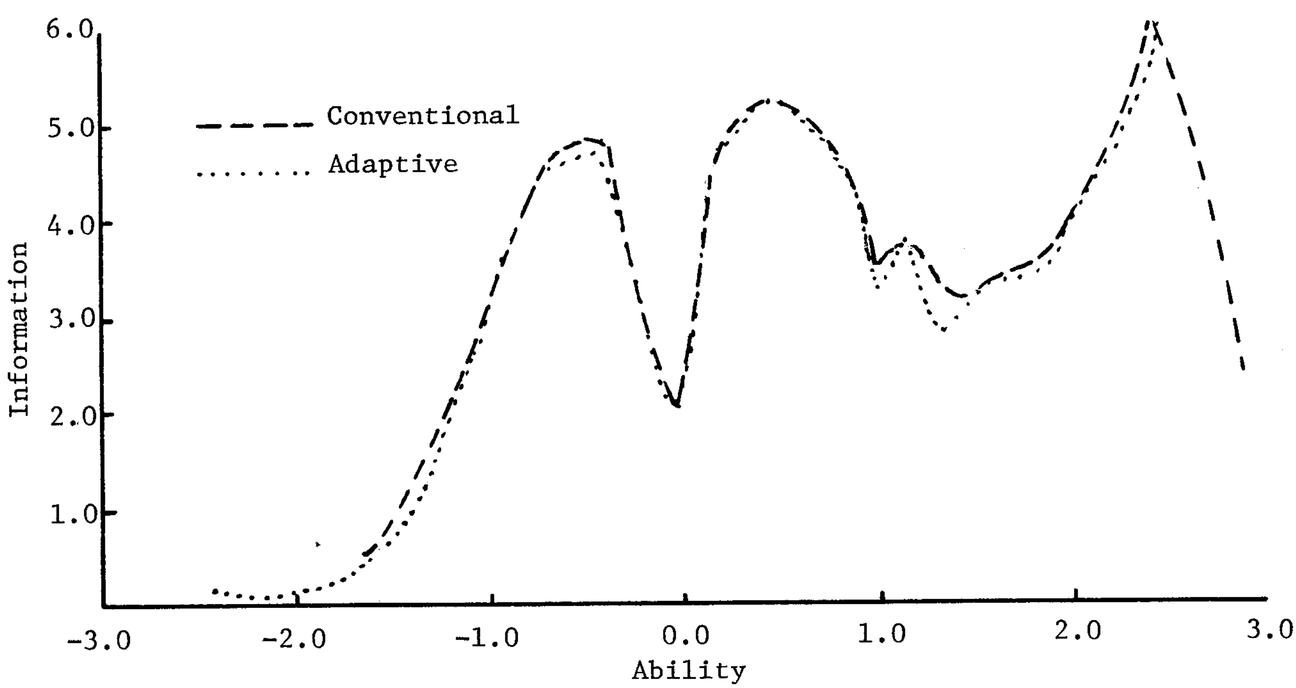


Figure 8
Observed Information Curves for Subtest 12 (Conventional Test = 18 Items; Adaptive Test = 6.23 Items Average, Range = 1 to 15 Items)



necessary to administer in the adaptive battery all of the items included in the conventional tests. Scores from the adaptive tests correlated highly with those from the conventional tests. Adaptive testing therefore, can reduce the time spent in testing; the time saved could then be used by the testees for other activities, such as additional instruction. It is also possible that adaptive achievement testing might have positive psychological advantages (e.g., Betz & Weiss, 1976), providing further beneficial effects on the psychometric characteristics of test scores. At the least, reduced testing time might result in more favorable attitudes of the testees toward the testing process.

In the adaptive testing strategy implemented in this study, test length is a direct function of the termination criterion employed. Testing was terminated within a subtest when none of the remaining items had a corresponding level of item information greater than .01 (an arbitrarily chosen value) at the testee's current estimated achievement level. More research is needed to determine optimal termination criteria.

The results of this study have also shown that the amount of information extracted by adaptive testing closely approximated that for conventional testing. That the information curves resulting from the adaptive and conventional strategies were found to be highly correspondent was to be expected from the way in which items were selected (based on item information) for the adaptive strategy. However, because of the inapplicability of maximum likelihood scoring in the early stages of item administration within a subtest, additional research is needed to develop and evaluate optimal procedures for item scoring. In addition, further research is needed for identification and evaluation of optimal procedures to order subtests for inter-subtest branching.

This study has demonstrated that an adaptive testing strategy, designed specifically for achievement test batteries, can substantially reduce the number of items administered in all subtests of the battery without reducing the precision of subtest scores. The strategy appears to be generalizable; it should be applicable to a variety of test batteries in which there is a fixed and relatively small subset of items for each subtest. Further research is needed to evaluate the performance of this adaptive testing strategy in other test batteries and in live testing situations. In addition, research is needed to modify the adaptive testing strategy to identify optimal procedures for the complete individualized administration of an achievement test battery.

References

Bejar, I. I., Weiss, D. J., & Gialluca, K. A. An information comparison of conventional and adaptive tests in the measurement of classroom achievement (Research Report 77-7). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1977.

Bejar, I. I., Weiss, D. J., & Kingsbury, G. G. Calibration of an item pool for the adaptive measurement of achievement (Research Report 77-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1977.

Betz, N. E., & Weiss, D. J. An empirical study of computer-administered two-stage ability testing (Research Report 73-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1973. (NTIS No. AD 768993)

Betz, N. E., & Weiss, D. J. Empirical and simulation studies of flexilevel ability testing (Research Report 75-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1975. (NTIS No. AD A013185)

Betz, N. E., & Weiss, D. J. Psychological effects of immediate knowledge of results and adaptive ability testing (Research Report 76-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1976. (NTIS No. AD A027170)

Betz, N. E., & Weiss, D. J. Simulation studies of two-stage ability testing (Research Report 74-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1974. (NTIS No. AD A001230)

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968, chap. 17-20.

Brown, J. M., & Weiss, D. J. An adaptive testing strategy for achievement test batteries (Research Report 77-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1977.

Gugel, J. F., Schmidt, F. L., & Urry, V. W. Effectiveness of the ancillary correction procedure. In C. L. Clark (Ed.), Proceedings of the first conference on computerized adaptive testing. Washington, DC: U.S. Civil Service Commission, 1976.

Larkin, K. C., & Weiss, D. J. An empirical comparison of two-stage and pyramidal adaptive ability testing (Research Report 75-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1975. (NTIS No. AD A006733)

Larkin, K. C., & Weiss, D. J. An empirical investigation of computer-administered pyramidal ability testing (Research Report 74-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1974. (NTIS No. AD 783553)

Lord, F. M. Discussion. In W. A. Gorham, J. F. Gugel, F. M. Lord, C. Jensema, F. L. Schmidt, & V. W. Urry. Computers and testing: Steps toward the inevitable conquest. Symposium presented at the 83rd Annual Convention of the American Psychological Association, Chicago, 1975. (PB-261-694) Washington, DC: U.S. Civil Service Commission. (NTIS No. AD 16851469).

Owen, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 1975, 70, 351-356.

Urry, V. W. A five-year quest: Is computer-assisted testing feasible? In C. L. Clark (Ed.). Proceedings of the first conference on computerized adaptive testing. Washington, DC: U.S. Civil Service Commission, 1976.

Urry, V. W. Tailored testing: A successful application of latent trait theory. Journal of Educational Measurement, 1977, 14, 181-196.

Vale, C. D., & Weiss, D. J. A simulation study of stradaptive ability testing (Research Report 75-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1975. (NTIS No. AD A020961)

Weiss, D. J. The stratified adaptive computerized ability test (Research Report 73-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1973. (NTIS No. AD 768376)

Weiss, D. J. Strategies of adaptive ability measurement (Research Report 74-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1974. (NTIS No. AD A004270)

Weiss, D. J., & Betz, N. E. Ability measurement: Conventional or adaptive? (Research Report 73-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1973. (NTIS No. AD 757788)

Acknowledgements

This research was supported by funds from the Army Research Institute, Air Force Human Resources Laboratory, Defense Advanced Research Projects Agency, Navy Personnel Research and Development Center and Office of Naval Research, and monitored by the Office of Naval Research under contract No. N00014-76-C-0627 NR150-389. Data utilized in this report were obtained from the Personnel and Training Evaluation Program (PTEP) of the Naval Guided Missile School at Dam Neck, Virginia. Appreciation is extended to Lieutenant Commander Lee J. Walker of PTEP and Dr. Myron A. Robinson of Data Design Laboratories at Norfolk, Virginia for their cooperation in this research.

Editor's Note

Although this paper was presented at the conference during this session, it was not on the final conference schedule. Since it was not available to the session discussant prior to the conference, discussion of the paper was not included in the discussant's comments.