AN EMPIRICAL COMPARISON OF TWO-STAGE AND PYRAMIDAL ADAPTIVE ABILITY TESTING

Kevin C. Larkin

and

David J. Weiss

Research Report 75-1

Psychometric Methods Program
Department of Psychology
University of Minnesota
Minneapolis, MN 55455

February 1975

Prepared under contract No. N00014-67-A-0113-0029
NR No. 150-343, with the Personnel and
Training Research Programs, Psychological Sciences Division
Office of Naval Research

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
Research Report 75-1

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
An Empirical Comparison of Two-stage and Pyramidal Adaptive Ability Testing

5. TYPE OF REPORT & PERIOD COVERED
Technical Report

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Kevin C. Larkin and David J. Weiss

8. CONTRACT OR GRANT NUMBER(s)
N00014-67-0113-0029

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Department of Psychology
University of Minnesota
Minneapolis, Minnesota 55455

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
P.E.:61153N PROJ.:RR042-04
T.A.:RR042-04-01
W.U.:NR150-343

11. CONTROLLING OFFICE NAME AND ADDRESS
Personnel and Training Research Programs
Office of Naval Research
Arlington, Virginia 22217

12. REPORT DATE
February 1975

13. NUMBER OF PAGES
27

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

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

18. SUPPLEMENTARY NOTES

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

testing
ability testing
computerized testing
adaptive testing
sequential testing
branched testing
individualized testing
tailored testing
programmed testing
response-contingent testing
automated testing

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

A fifteen-stage pyramidal test and a 40-item two-stage test were constructed and administered by computer to 111 college undergraduates. The two-stage test was found to utilize a smaller proportion of its potential score range than the pyramidal test. Score distributions for both tests were positively skewed but not significantly different from the normal distribution. The pyramidal test's score distributions tended to be platykurtic while the two-stage test's distribution tended to be leptokurtic. The

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

assignment of subjects to measurement subtests in the two-stage test was more accurate than in a previous empirical investigation since the misclassification rate was less than 1%. Comparison of scoring methods for the pyramidal strategy supported earlier findings that the average difficulty scoring methods were most useful. The correlations between scores on the two adaptive strategies ranged from $r=.79$ to .84. Both adaptive strategies appeared to adapt item difficulties to individual differences in abilities so as to reduce chance effects due to guessing. The pyramidal strategy seemed to be slightly more successful in eliminating guessing than the two-stage strategy. Results are discussed with respect to internal consistency reliabilities, stabilities, and the relation of each strategy to conventional testing. Simulation studies are suggested to further delineate the optimum characteristics of each testing strategy.

Contents

AN EMPIRICAL COMPARISON OF TWO-STAGE AND PYRAMIDAL ADAPTIVE ABILITY TESTING

The administration of ability test items by means of an interactive computer system has enabled test administrators to tailor or adapt tests to individual differences in testee ability. Items are selected by a set of rules or "strategy" determined prior to testing (see Weiss, 1974, for a discussion of the various adaptive testing strategies). At one or more points in the testing, a testee's responses to previously administered items are evaluated, and a tentative estimate of ability is made. Subsequent items are generally selected so that their difficulties are close to the testee's estimated ability. This procedure permits testing time to be shortened in comparison to conventional paper and pencil methods of testing without reducing either the reliability or validity of the test. Computerized adaptive testing also has other advantages over conventional tests (see Weiss and Betz, 1973).

Empirical research on two adaptive strategies, the pyramidal test and the two-stage test, has been reported in the present series of research papers (Betz & Weiss, 1973; Larkin & Weiss, 1974). In both of these studies, the adaptive test was compared to a conventional test on a number of psychometric criteria. The present study directly compares the two adaptive strategies using the same group of subjects.

Pyramidal Tests

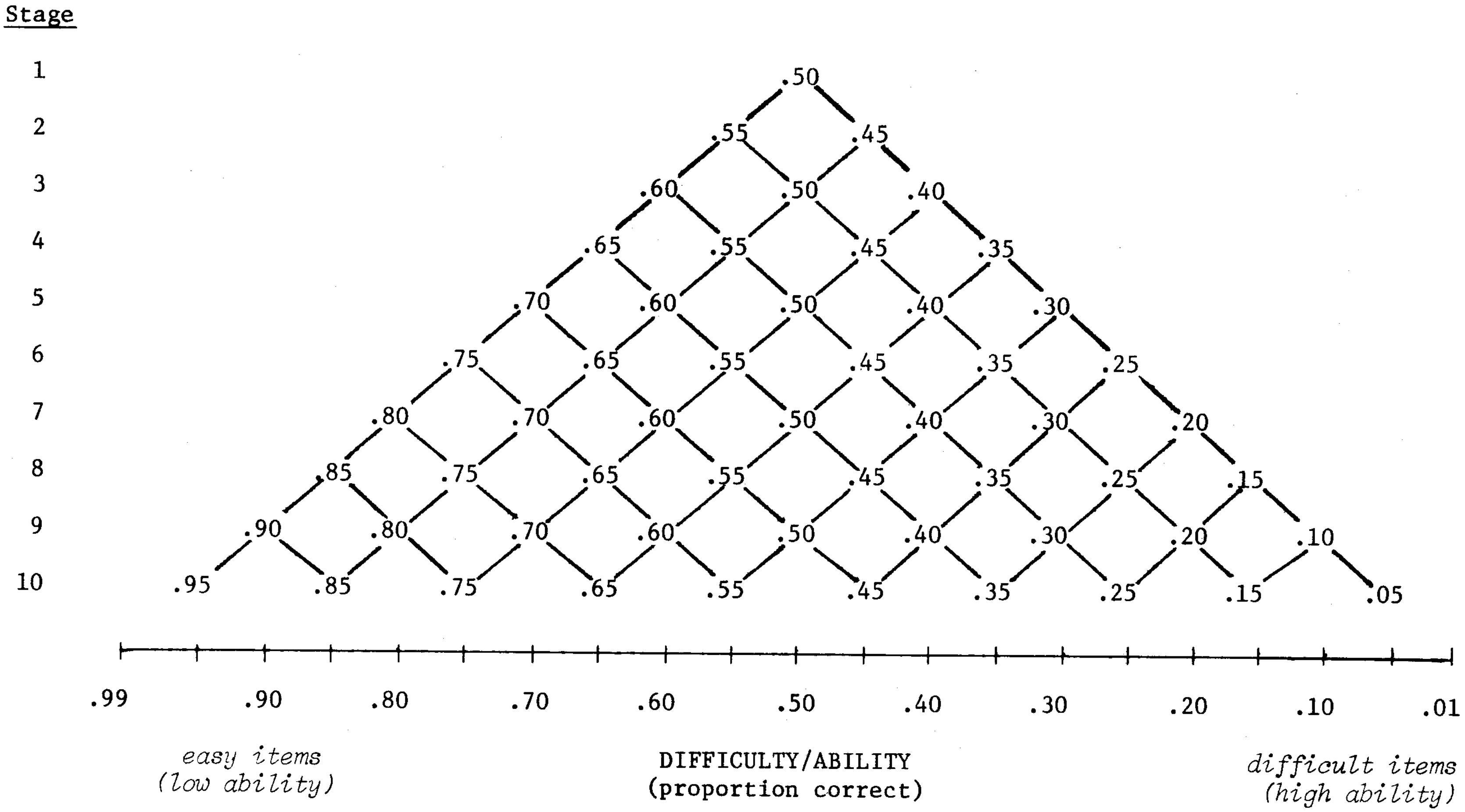
The pyramidal testing method structures items into a triangular configuration according to item difficulties. Item administration follows the general branching rule that a more difficult item follows a correct response while an easier item follows an incorrect response. Figure 1 illustrates a typical pyramidal test. The first item administered is at the top of the pyramidal structure and is usually one of median difficulty (proportion correct, $p=.50$) based on previous item analyses. The second item administered to any testee depends on whether his/her response to the first item is correct or incorrect. If the testee answers the first item correctly, a more difficult item $(p=.45)$ is presented next. An item of lesser difficulty $(p=.55)$ is presented next if the initial item is answered incorrectly. Thus, there are two items available at the second level or "stage" of the pyramid. Branching to the third stage depends on the correctness of the response to the second-stage item. This process is repeated until the testee has attempted one item at each of a fixed number of stages.

The increment in difficulty following a correct response in Figure 1 is equal to the decrement in difficulty following an incorrect response. Thus, branching within this pyramidal structure uses an "equal offset." Unequal offsets with smaller increments than decrements can be used as a correction for guessing (Weiss, 1974, p. 16).

The number of items to be answered by any testee is small when compared to the total number of items in the pyramidal structure. In general $[n(n-1)]/2$ items are needed to construct a pyramid of n stages when one item is attempted at each stage.

Figure 1

Item difficulties in a ten-stage pyramidal test structure



Many ways of scoring pyramidal tests have been developed (see Weiss, 1974, pp. 30-34). The ranked difficulty of the final item has been used as the individual's score (Bayroff, Thomas, & Anderson, 1960; Seeley, Morton & Anderson, 1962; Waters & Bayroff, 1971). Testees completing the pyramid shown in Figure 1 could receive scores of from 1 to 1^ under this scoring method, since there are only ten items available at the tenth stage of testing. The number of rank positions can be increased by assigning a higher rank to those subjects answering the final item correctly than to those who do not (Bayroff & Seeley, 1967; Waters, 1964). The difficulty of the final item attempted has also been used to estimate an individual's ability (Bayroff, 1969). Another scoring method branches the testee to a hypothetical *(n+1)th* item following the final item and estimates its difficulty (Hansen, 1969; Lord, 1971b; Weiss, 1974, p. 31). The difficulties of all items attempted or all items correctly answered may be averaged to provide a score based on more information (Larkin & Weiss, 1974). Lord (1970, 1971b) has recommended an averaging method which excludes the first item (since all testees attempt it) and includes the *(n+1)th* item. Hansen (1969) has proposed a more complex scoring method which assigns an estimated score to each item in the pyramid, whether or not it is attempted.

Weiss (1974) compares pyramidal tests with other strategies of adaptive testing. The research literature on pyramidal adaptive tests has been reviewed by Weiss and Betz (1973) and summarized by Larkin and Weiss (1974).

Two-stage Tests

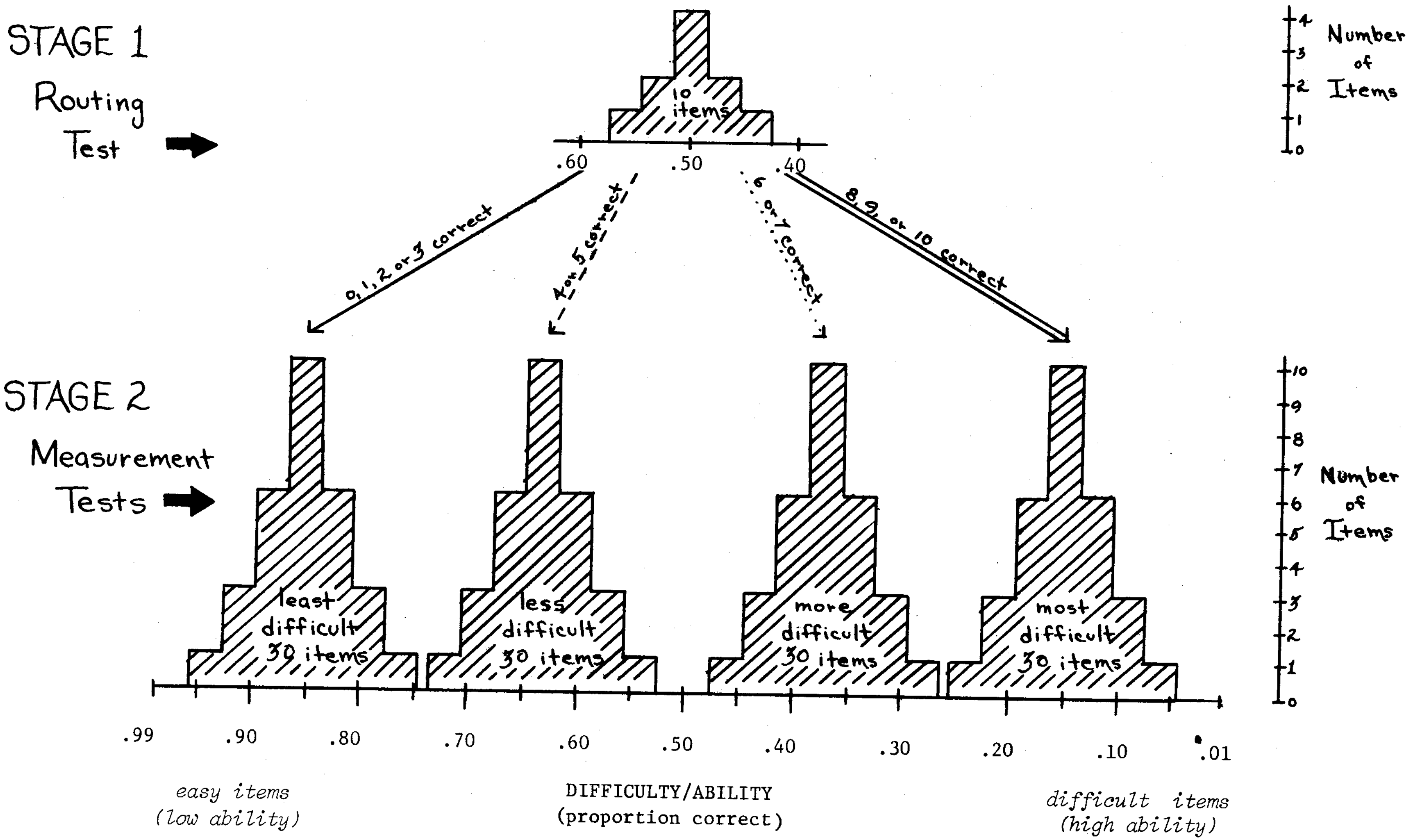
A two-stage test consists of a preliminary or routing test followed by one of several measurement tests. Figure 2 illustrates a sample two-stage structure. The purpose of the routing test is to provide an approximate estimate of the testee's ability level so that a measurement test of appropriate difficulty can be selected for each testee. The routing test can be composed of items with difficulties either peaked at the ability level of the group taking the test (as shown in Figure 2) or distributed throughout the range of ability under consideration (see Weiss, 1974, pp. 4-7). The measurement tests are usually peaked tests of differing levels of difficulty. The routing test is administered to the testee and his/her score is determined. A measurement test of appropriate difficulty is selected, based on the testee's score on the routing test. The measurement test is then administered and the testee's score is determined.

Variants of the two-stage routing procedure (see Weiss, 1974, p. 7) include double-routing and "sequential" procedures (Cleary, Linn, & Rock 1968a,b; Linn, Rock, & Cleary, 1969). The former requires two routing tests to be administered. A testee's score on a preliminary routing test determines which of several intermediate routing tests are attempted. Branching to an appropriate measurement test is based on the testee's performance on the second routing test. The sequential procedure involves computing likelihood ratios after each response to items in the routing test. Branching to a measurement test occurs when the likelihood ratio permits a classification of the individual.

Most methods of scoring two-stage tests have used information from

Figure 2

A Two-Stage Strategy with Peaked Routing Test



both the routing and measurement subtests. Lord (1971c) and Betz and Weiss (1973) have combined maximum likelihood ability estimates from the routing and measurement subtests to determine an overall estimation of a testee's ability. Linn, Rock, & Cleary (1969), on the other hand, did not include a testee's performance on the routing test in some of their scoring procedures.

Weiss (1974) compares two-stage tests with other strategies of adaptive testing, and discusses potential advantages and limitations of this approach. Research literature on two-stage adaptive testing has been reviewed by Weiss & Betz (1973) and Betz & Weiss (1973).

Research Comparing Two-stage and Pyramidal Tests

The only study including both two-stage and pyramidal testing strategies was reported by Linn, Rock & Cleary (1969). That study, using "real-data simulation" methods, was based on the responses of a large group of testees to a 190-item conventional paper-and-pencil test. The item responses were then used to simulate a testee's responses to two-stage and pyramidal adaptive testing strategies. Five different two-stage strategies were compared to two pyramidal strategies.

The first two-stage test included a 20-item routing test with a rectangular distribution over a "broad range" of item difficulties, and four 20-item measurement tests. The second employed a double-routing procedure. A testee's score on a 10-item routing test determined which of two second-stage 10-item routing tests was administered. Scores on the second routing test branched the testee to one of four 20-item measurement tests. The third two-stage procedure used a 20-item "group discrimination" routing test. Items in that test were those which showed the largest differences in proportion correct between groups divided into quartiles on total scores for the original 190 items. The routing test in the two final strategies involved computing likelihood ratios after each item, and branching occurred when the likelihood ratio permitted a classification of the individual into groups based on scores derived from the parent 190 items. These methods were called "sequential" procedures. Both a three-group and a four-group sequential approach were used. Linn et al. used two methods to score their two-stage tests. One used the information obtained from the routing test while the other did not.

Linn et al. studied two variations of the pyramidal strategy. The first pyramidal test had ten stages with an entry point of $p=.65$, a step size of .02 and an equal offset. Items were weighted according to difficulty, and scores represented the sum of the weights of items attempted by each testee. The second pyramidal strategy consisted of five stages with five items per stage (see, e.g., Weiss, 1974, pp. 25-26). Branching occurred from block to block. This pyramid was scored using a weighted scoring scheme similar to that used for the single-item pyramid.

All seven adaptive tests were compared to five conventional subtests of from 10 to 50 items selected from the same 190-item parent test. Scores on the two-stage strategies correlated from .93 to .97 with scores on the 190-item parent test, while the shortened conventional tests had correlations of from .89 to .96 with the full conventional test. The 25-item pyramid showed a comparable correlation (.95), but the ten-item pyramid correlated only .87 with the parent test.

Since all the items in the adaptive and shortened conventional tests were also included in the longer parent test, and since the correlations with the parent test increased as the length of the shorter tests increased, it is possible that the degree of correlation obtained in this study could be due partly to the number of items in common between the tests.

When achievement test criteria were obtained, scores on the ten-stage pyramidal test correlated higher with the criterion measures than did scores on conventional tests of the same length in seven of eight comparisons. The 25-item pyramids correlated more highly with the criteria than the 50-item conventional tests. With one exception, the two-stage tests also achieved higher correlations with the criterion achievement tests than did conventional tests of the same length. Under one scoring procedure the 40-item group discrimination two-stage test was more highly correlated with outside criteria, in four of eight comparisons, than even the 190-item parent conventional test.

Linn et al.'s data permit a comparison of the relative validity of their two-stage and pyramidal tests as predictors of the achievement test criteria. Their data show that, with the exception of the sequential two-stage strategy, two-stage tests had higher correlations with the criteria than did the pyramidal tests. The ten-item pyramidal test had the lowest validities of all of the adaptive tests, and the validities of the 25-item block branching pyramid were about equal to those of the sequential two-stage test. Within the two-stage tests, the group discrimination approach had slightly higher validities than the other two-stage tests.

These comparisons of the relative validity of the two-stage and pyramidal strategies did not take account of the relative numbers of items in the different tests. While the two-stage tests were all composed of about 40 items, only 10 items were administered in one pyramidal test and 25 in the other. Linn et al. (pp. 142-143) estimated the lengths of conventional tests parallel to the 190-item parent test which would be necessary to achieve the same validity as each of the adaptive tests. When these values were compared to the actual adaptive test lengths, an index of "relative saving in test length" was obtained. The group discrimination and three-group sequential methods showed the highest ratios, followed by the 25-item and 10-item pyramidal strategies. The four-group sequential method showed the lowest ratios.

Purpose

Although Linn et al. (1969) used both pyramidal and two-stage tests in their study simulating adaptive testing, their major objective was to study the relationships between short adaptive and conventional tests and longer parent tests or achievement test criteria. The present investigation is one of a series of studies designed to further compare adaptive testing strategies using other criteria. These studies use actual computer administration of adaptive tests to groups of college students. The results of different adaptive testing strategies have been compared with those obtained from conventional testing approaches (Betz & Weiss, 1973; Larkin & Weiss, 1974) with respect to the accuracy of ability estimation, test-retest stability, internal consistency reliabilities, and other psychometric characteristics. In addition, more fundamental questions about each strategy are under consideration, including the investigation of various item difficulty structures for each of the adaptive strategies, problems in determining branching or routing rules, and the determination of meaningful and reliable scoring methods for each adaptive strategy.

In this series of studies, all tests, both conventional and adaptive, were constructed for administration by computer (DeWitt & Weiss, 1974). Testing strategies were administered two at a time so that scores from one adaptive strategy could be compared with those from another, and so that scores from adaptive and conventional tests could be directly compared. In order to determine the stability of scores from each of a number of scoring methods, each testee was administered the same test on two occasions with periods averaging about six weeks between the initial and final testing. In some studies, conventional and adaptive strategies were paired on both test and retest and the comparative stabilities of the two strategies were studied. Other studies focused on comparisons among the various adaptive strategies.

The present analysis was undertaken for the purpose of directly comparing the psychometric characteristics of scores obtained from a two-stage strategy and the pyramidal approach. Previous studies in this series have reported the results of analyses of computer-administered two-stage (Betz & Weiss, 1973) and pyramidal tests (Larkin & Weiss, 1974) in comparison with conventional tests. However, a different group of subjects was used in each of those studies. In the present study, the characteristics of scores derived from two-stage and pyramidal tests are compared directly using the same group of subjects.

METHOD

One set of test data was derived from the administration of a two-stage test and a pyramidal test to 111 subjects.

The 15-stage pyramidal item structure was composed of 120 items. Each testee completed only fifteen items. The two-stage test required 130 items for its construction and each subject answered 40 items. Both tests drew items from the same item pool, and eighty items were common to both test structures. Although each testee could be administered a maximum of 15 items common to both tests, it was also possible that a testee could receive no common items.

In order to detect the presence of the effects of boredom or fatigue, the order of presentation was randomized on both testings. Each adaptive test was administered first to half the testees and administered second to the remaining testees.

Test Construction

Item Pool

The item pool was composed of 369 five-alternative multiple-choice vocabulary questions normed on college undergraduates (McBride & Weiss, 1974).

Using estimates of item difficulty (proportion correct) and item discrimination (biserial correlation with total score on the norming tests) approximations to the normal ogive item parameters *a* and *b* (Lord & Novick, 1968, pp. 376-379) were determined using the following formulas:

$$a = \frac{r_b}{\sqrt{1-r_b^2}} \qquad (1)$$

$$b = \frac{-\sqrt{1+a^2}}{a} \cdot \Phi^{-1} \qquad (2)$$

where a is the normal ogive index of discrimination
b is the normal ogive index of difficulty
r_b is the biserial correlation of item response and total score

and Φ^{-1} is the inverse of the cumulative normal distribution corresponding to the proportion correct.

Items with biserials lower than .30 were not used in the item pool. The norming studies indicated that there was some difficulty-discrimination interaction such that the pool contained disproportionately more highly discriminating items in the lower range of item difficulty.

Construction of the Two-stage Test

The two-stage test used in this study was composed of a 10-item routing test and four 30-item measurement tests. This adaptive test was the "Two-stage 2" test in a simulation study in a previous report in this series (Betz & Weiss, 1974).

Routing test. In order to make a good initial assessment of ability and to assign testees to measurement tests while minimizing the probability of an assignment error, the 10 items in the routing test were selected to have a high mean discrimination. As shown in Table 1, mean discrimination for the routing test was a=.702. The standard deviation of the item discriminations was .163. Appendix A, which shows difficulty and discrimination values for each item in the routing subtest, indicates that the lowest discrimination was a=.50 and the highest was a=.98.

The routing subtest was a peaked test of median difficulty items which were highly discriminating. The items in the routing subtest had a mean difficulty level of b=-.232. Table 1 shows that the standard deviation of the item difficulties in the routing test (.50) was very low when compared to those of the measurement tests.

After the routing test was completed, an estimate of the testee's ability was made in standard units (see Betz & Weiss, 1974, pp. 11-12). Subjects were assigned to the measurement test closest in difficulty to their estimated ability. Thus, those testees with from 0 to 4 items correct on the routing test were assigned to the least difficult of the measurement tests. Those with scores of 5-6, 7-8, and 9-10 were routed to one of the three more difficult measurement tests.

Table 1

Means and Standard Deviations of Normal Ogive Item Parameters for Two-stage and Pyramidal Tests

Test	No. of Items	Difficulty (b)		Discrimination (a)	
		Mean	S.D.	Mean	S.D.
Two-stage (all items)	130	-.072	1.251	.633	.183
Routing	10	-.232	.050	.702	.163
Measurement 1	30	1.725	.558	.530	.126
Measurement 2	30	.350	.297	.684	.214
Measurement 3	30	-.709	.189	.611	.122
Measurement 4	30	-1.603	.373	.683	.213
Pyramid	120	-.094	1.256	.799	.457

Measurement tests. In selecting items for each of the four measurement tests, the following rationale was used. The quantity $\underline{a}(\overline{b}_i-\overline{b}_r)$ was computed, where $\overline{b}_r$ is the mean difficulty of the routing test. The $\underline{a}$ parameter is the mean discrimination for all 130 items in the two-stage structure, i.e., .633; $\overline{b}_i$ represents the mean difficulty of the measurement test in question. Betz and Weiss (1974) have shown that to obtain four measurement tests suitable for subjects routed to them, the values required for $\overline{a}(\overline{b}_i-\overline{b}_r)$ were 1.239, .368, -3.02 and -.868.

Table 1 shows that the average of the discrimination parameters for the measurement tests ranged from .530 to .684, and that the average variability of discrimination values for the measurement tests was about the same as average variability of item discriminations in the routing test. Measurement tests were not as peaked as the routing test as indicated by the larger ranges and standard deviations of their difficulties. The average difficulties of the measurement tests, as shown in Table 1, approximated the desired values, but measurement test 1 was somewhat more difficult than the target value, and measurement tests 3 and 4 were somewhat easier. Appendix A gives the normal ogive item parameters for the items in each measurement test.

Scoring. The two-stage test was scored by the same method used by Betz and Weiss (1973, 1974) who adapted their method from studies by Lord (1971c). Essentially, maximum likelihood estimates of ability were obtained from both subtests and then weighted and summed. The measurement test was given three times the weight of the routing test because there were three times as many items in it as in the routing test.

The formula used to obtain the ability estimates for both subtests completed by each testee was:

$$\hat{\Theta} = \frac{1}{\bar{a}_i} \Phi^{-1} \left[\frac{(x/m)-c}{1-c}\right] + \bar{b}_i \qquad (3)$$

where $\bar{a}_i$ is the mean discrimination of the subtest

x is the number correct

m is the number of items in the subtest

c is the chance score level

$\bar{b}_i$ is the mean difficulty of items in the subtest

and Φ^{-1} is the inverse of the cumulative normal distribution function corresponding to the proportion correct.

For perfect scores $(x=m)$, Θ could not be determined. Therefore, when x was equal to m, it was replaced by $x=m-.5$. For scores at or below chance $(x \leq cm)$, $\hat{\Theta}$ was also indeterminate and x was replaced by $x=cm+.5$.

The scores of the subtests were combined in the following way:

$$\hat{\Theta} = \frac{\hat{\Theta}_1 + 3\hat{\Theta}_2}{4} \qquad (4)$$

where $\hat{\Theta}$ is the combined ability estimate

$\hat{\Theta}_1$ is the ability estimate obtained from the routing test

$\hat{\Theta}_2$ is the ability estimate obtained from the measurement test.

This combined ability estimate can be interpreted as a standard normal deviate (see Betz & Weiss, 1973, pp. 14-15).

Construction of the Pyramidal Test

The pyramidal test used in this study was Pyramid 3, studied by Larkin & Weiss (1974). It was composed of fifteen stages with a constant step size. An up-one/down-one branching rule was used. Since $n(n+1)/2$ items are needed for the construction of an n-stage pyramid, $15(15+1)/2$ or 120 items were selected from the item pool. The initial item was of median difficulty for the testees of the norm group. The step size, that is the increment or decrement in item difficulty from one stage to the next, had a mean value of b=.199, and a standard deviation of .08.

After establishing the initial item difficulty and step size, the available items in the pool were divided into 29 groups on the basis of difficulty. All items in a group had about the same b value and an a value of at least .30. The items required were selected from each group according to their discriminations. The items with the highest discriminations in each group were selected for use in the pyramidal test. Paterson (1962) has suggested that items in a pyramidal test be ordered within each column according to discrimination with the most discriminating items appearing first. This suggestion was followed in construction of this pyramidal test, as shown in Appendix B which gives the normal ogive difficulty and discrimination estimates for each item in the pyramidal test. The item difficulties ranged from b=-2.86 to b=2.61. The discrimination values varied from a=.41 to a=3.00.

Appendix B indicates that the initial item, which was presented to all testees, had a difficulty of b=-.05. If the subject answered this item correctly, he/she was branched to a more difficult item (b=.14) at stage 2. An incorrect response branched the testee to an item easier (b=-.21) than the initial item. The branching process continued until each testee had attempted 15 items.

The means and standard deviations for difficulty and discrimination are shown in Table 1. The average difficulty of the items in the pyramidal structure was b=-.094, with a standard deviation of 1.256. The average discrimination of the pyramid items was a=.799. When all items in each adaptive test were considered, Table 1 shows that the overall difficulties were almost the same. The 120 items in the pyramidal structure and the 130 items in the two-stage test had very similar means and standard deviations of item difficulties. However, the pyramid was composed of more highly discriminating items and the variance of the item discriminations was much higher in the pyramidal test.

Scoring. In order to compare ability estimates derived from various scoring methods, four different methods were used to estimate ability. These four methods were among those used in a previous investigation of pyramidal testing (Larkin & Weiss, 1974). Method 1 was the number of correct responses. This has been the most common scoring method used in other studies. For a pyramid of 15 stages, 16 different number correct scores are possible (0 to 15). Method 2 was the mean difficulty of the items attempted by each testee. An approach similar to this involves averaging the difficulties of all items but the first (since every testee attempts it) and including a hypothetical sixteenth item (Lord, 1970, 1971b). Method 3 averages the difficulties of the correctly answered items only. Under method 4, subjects were scored by the difficulty of the final item attempted in the pyramid; since the branching strategy actually adapts the difficulties of the items to the ability of the testee, the difficulty of the final item reached should reflect the testee's ability level (assuming that the pyramidal structure has enough stages). Two other scoring methods, the *(n+1)th* difficulty score and the all-item score (Hansen, 1969) were found in previous research (Larkin & Weiss, 1974) to correlate perfectly with the number correct score and mean difficulty of all items attempted respectively. Consequently, these two scoring methods were not used in the present analyses.

Test Administration and Subjects

Cathode ray terminals (CRT's) acoustically coupled to a time-shared computer systerm were used to administer both the two-stage and pyramidal test (DeWitt & Weiss, 1974). Items were presented one at a time on the CRT screen; subjects responded by typing a number corresponding to the correct alternative to each multiple-choice item. A total of 55 items (15 from the pyramidal test and 40 from the two-stage test) was administered to each testee. The order of presentation of the tests was randomized over subjects. Fifty-six testees completed the pyramidal test first and the 55 remaining testees completed the two-stage test first. Subjects were informed at the completion of testing of the total number of items they answered correctly.

The testees were undergraduates enrolled in general psychology or psychological statistics courses at the University of Minnesota. Because this combination of adaptive tests was given as the second session of a two-part study, all had had previous experience with computer-administered tests. All subjects were given the opportunity to review instructions explaining the operation of the CRT's prior to testing. A proctor was available in the testing room to begin the testing and to provide further assistance to any testee having difficulty with the equipment. No time limit was imposed. Testees were informed that they might take as much time as necessary to finish the tests.

Analysis

The data analyzed in the present study consisted of five scores, one two-stage score and four pyramidal scores, for each testee.

Order Effects

The effects of the order of administration on test scores were investigated by comparing scores of the testees who received each strategy first with those who received that strategy second in the series of two tests. In this manner fatigue, practice, or carry-over effects between strategies could be detected. Because the scores were expected to be highly correlated a one-way multivariate analysis of variance was used with all five scores simultaneously considered as dependent variables.

Characteristics of Score Distribution

One objective of the present study was to compare the distributions of scores on the 40-item two-stage test with those obtained from each method of scoring the 15-stage pyramidal test. The appropriateness of the test difficulty, the relative variabilities of each scoring procedure, and the shape of the obtained score distributions were examined.

Because different units were used in scoring the tests, the standard deviation of each scoring method was divided by the potential range of scores under that method. The resulting value is an index of relative variability (Betz & Weiss, 1973). This index shows the effective utilization of the entire score range for each scoring method. The range of possible scores on the two-stage test was derived using Formulas 3 and 4 to compute estimates of $\hat{\theta}$ for perfect and chance scores. This range was 4.66-(-5.30)=9.96. The ranges for the pyramidal scoring methods were as follows: 1) The number correct range was 15; 2) The range for the mean-difficulty-attempted score was the difference between the score made by a testee answering all items correctly and the score of one responding incorrectly to all items. This value was 2.79; 3) The range for the mean-difficulty-correct score was the difference between the score of a subject with 15 correct responses and the lowest $(n+1)th$ score. The latter value was used since a testee with no items answered correctly would have a mean-difficulty-correct score which was undefined. This range was 4.42; 4) The final item difficulty range was the difference between the easiest and most difficulty terminal items, or 5.48.

In addition to the mean and variability indices, the skew and kurtosis of each distribution was computed and the significance and direction of its departure from normality were determined (McNemar, 1969, pp. 25-28, 87-88).

Relationships between Two-stage and Pyramidal Scores

To determine the relationships among the pyramidal scores and their relationships to two-stage scores, product-moment correlations and correlation ratios (eta) were computed. The latter were computed to determine whether the relationships between scores on the two strategies were curvilinear. In determining the etas, both the regression of two-stage scores on pyramidal scores and the regression of pyramidal scores on two-stage scores were computed.

Internal Consistency Reliability

Data on the reliabilities of the two-stage and pyramidal tests are important to provide a point of reference for interpreting the correlations between scores on the two adaptive strategies.

The internal consistency reliability of the two-stage test was determined by Hoyt's (1941) method. This index can be computed only if every subject attempts each item on a test. For this reason, the two-stage test had to be treated as five separate tests. Reliabilities were computed separately for the routing test, using the responses of the total group of subjects, and for each of the four measurement tests, using the responses of those subjects routed to each measurement subtest. To compare the internal consistencies of the 10-item routing test with that of 30-item measurement tests, the Spearman-Brown formula was used to estimate the reliability of a 30-item routing subtest based on the testees' responses to 10 items.

Because all testees do not answer the same subset of items under the pyramidal strategy, its internal consistency reliability cannot be determined satisfactorily (see Larkin & Weiss, 1974). Consequently, to make meaningful comparisons between the reliabilities of the pyramidal and two-stage tests, the test-retest correlations for each strategy determined from two previous empirical studies (Betz & Weiss, 1973; Larkin & Weiss, 1974) were used.

Mis-routing

Mis-routing occurs in the two-stage strategy when a testee is routed to measurement tests of inappropriate difficulty. The following criteria were used (see Betz & Weiss, 1973) to determine the proportion of testees who were mis-routed. All testees who obtained perfect scores (30) on their measurement subtest were considered to have been routed to a test too easy for them. Those testees with subtest scores at or below chance level (i.e., 6 correct) were considered to have been assigned to a measurement test too difficult for them. If a testee met either of the two criteria, he/she was classified as having been mis-routed by the routing test.

Intercorrelations of Pyramidal Scores

Product-moment correlations were computed for all pairs of pyramidal scoring methods to determine the interrelationships among them. Correlation ratios were computed and compared with the product-moment correlations to detect the presence of possible curvilinear relationships.

RESULTS

Order Effects

Table 2 shows the means and standard deviations by scoring method and strategy for the groups completing pyramidal or two-stage tests first. The one-way multivariate analysis of variance resulted in an F-value of .92 with an associated probability of .47. Thus the two sets of mean scores obtained under the two orders of administration were not significantly different. As a result, the data from both order groups were combined for all further analyses.

Table 2

Means and Standard Deviations for Subgroups Completing Pyramidal and Two-Stage Tests in Different Orders

	Pyramid First (N=56)		Two-stage First (N=55)	
Test and Scoring Method	Mean	S.D.	Mean	S.D.
Pyramidal Test				
Number Correct	8.21	2.55	7.64	2.24
Mean difficulty--attempted	0.10	0.56	-0.09	0.53
Mean difficulty--correct	-0.02	0.61	-0.22	0.57
Difficulty of final item	0.17	0.97	-0.06	0.88
Two-stage Test	-0.16	1.39	-0.50	1.19

Score Distributions

Pyramidal test. Descriptive statistics for the pyramidal and two-stage test scores are presented in Table 3. The mean number correct score of 7.93 indicated that the subject group as a whole answered approximately half the 15 items in the pyramid correctly, suggesting that the difficulty of the test was appropriate for the ability of the group tested. The two mean difficulty scoring methods and the final item difficulty scoring methods all had means of about 0.0. Since the test was composed of items with a mean difficulty of -.094, this result was expected. These results also suggest that there were few items answered correctly as a result of guessing, on the average, since guessing would have resulted in scores above the average of the norming group.

Table 3

Descriptive Statistics for Distributions of Scores from Pyramidal and Two-stage Tests (N = 111)

Test and Scoring Method	Mean	Median	S.D.	Proportion of Range Utilized	Skew	Kurtosis
Pyramidal Test						
Number Correct	7.93	7.43	2.41	.16	0.58*	0.08
Mean difficulty--attempted	0.01	-0.12	0.55	.20	0.42	-0.47
Mean difficulty--correct	-0.12	-0.23	0.60	.14	0.03	0.19
Difficulty of final item	0.06	-0.08	0.93	.17	0.44	-0.20
Two-stage Test	-0.33	-0.54	1.30	.13	0.35	0.29

*Statistically significant at p<.05.

The variabilities for each scoring method are also shown in Table 3. The final item difficulty score had a standard deviation of about 1.0, again reflecting the characteristics of the standardized b-values. Because of the restriction in the range of possible score values resulting from the use of averages, the two mean difficulty scores, also computed from b-values, had standard deviations only about half as large as the final item difficulty scoring method. When variability is expressed as a proportion of each method's potential range, as shown in Table 3, the scoring methods are more easily compared. The mean-difficulty-correct scoring method utilized the smallest proportion of its available range (.14), while the mean-difficulty-attempted method used the largest proportion of its range (.20). The number correct score and the final item difficulty scores both utilized about the same proportion of their range (.16 and .17).

All scoring methods had distributions which were slightly positively skewed. The distribution of number correct scores was the most highly skewed, and its skewness was significantly different from zero skew. The difficulty of all items correctly answered showed almost no skew.

Two score distributions--mean-difficulty-attempted and difficulty of final item--were platykurtic, although not significantly so. The number correct and mean-difficulty-correct distributions were slightly leptokurtic. The flattest distribution was that of the mean-difficulty-attempted scoring method. When both skewness and kurtosis are considered, the mean-difficulty-correct scores showed least departure from a normal distribution.

Two-stage test. The two-stage test scores, expressed in standard units, had a mean of -0.33 and a standard deviation of 1.30. This mean was slightly lower than that observed in the standardized pyramidal scores. The two-stage test utilized a smaller proportion of its possible range (.13) than any method of scoring the pyramidal test.

The distribution of two-stage scores was slightly positively skewed and was slightly leptokurtic, although in neither case was it significantly different from a normal distribution. The skewness was comparable to that of most methods of scoring the pyramidal test, but the kurtosis indicated that the two-stage score distribution was more peaked than those of the pyramidal test.

Table 4 summarizes the performance of the total group of testees on the 10-item routing test.

Table 4

Means and Standard Deviations of Scores on Subtests of the Two-Stage Test

		Subtest				Composite	
		Routing Test		Measurement Test		Two-stage Score	
		(Number Correct)		(Number Correct)		(Standard Score)	
Subject Group	N	Mean	S.D.	Mean	S.D.	Mean	S.D.
All Subjects	111	5.58	2.61	18.56	5.04	-0.33	1.30
Assigned to Measurement test 1	21	9.33	0.48	17.00	6.01	1.48	0.97
Assigned to Measurement Test 2	20	7.40	0.50	17.20	4.43	0.22	0.65
Assigned to Measurement Test 3	27	5.63	0.49	18.59	4.87	-0.54	0.70
Assigned to Measurement Test 4	43	2.86	1.15	19.93	4.66	-1.33	0.81

Also shown are descriptive statistics of scores for the testees assigned to each measurement subtest. On the routing test, the mean number of items correct over all subjects was 5.58 out of 10 items, suggesting that the routing test was peaked at a difficulty appropriate for the group taking the test. That is, item difficulties for the group tested averaged about .60 which is the expected median difficulty after chance has been taken into account. The standard deviation of number correct scores was relatively large (2.61), indicating that the routing test was effective in making an initial separation of testees according to ability. The mean number correct across all four measurement tests (18.56) showed that after testees had been routed into the measurement test, they answered slightly more than half the measurement test items correctly. For each 30-item measurement test considered separately, the mean number correct varied from 17.00 to 19.93 (or between 57 and 66 percent correct). These findings imply that the measurement tests were also of appropriate difficulty for the groups of testees routed to them. These results, however, suggest that there were somewhat more successes due to guessing in the two-stage test than in the pyramidal test.

The variability of scores for each of the four subject groups was relatively constant in three of the measurement tests; measurement test 1 had a slightly larger variability of scores than the other measurement tests. The variability in routing test scores for those subjects assigned to the least difficulty measurement test (1.15) was larger than that for the other groups due solely to the specifications of the routing procedure (i.e., a larger range of routing scores led to the assignment of testees to measurement test 4, the least difficult measurement test).

Relationship between Two-stage and Pyramidal Scores

Eighty items were common to both the pyramidal and two-stage item pools. The number of times a testee was administered the same item twice (once under each strategy) ranged from 0 to 13 with a mean of 6.02 and a standard deviation of 3.51. The correlations between the two tests are thus likely to be somewhat inflated due to the tendency of subjects to make the same responses to an item in both the two-stage and pyramidal test, and should be interpreted with caution.

Table 5 shows the results of the regression analysis of the relationship between scores on the two-stage test and scores on the pyramidal test. Product-moment correlations ranged from .79 for the mean-difficulty-correct scoring method to .84 for the number correct scoring method, Correlation ratios ranged from .83 to .88. There was no general tendency toward curvilinear relationships. In only one of the regressions was curvilinearity significant to the .05 level. Thus, the relationship between scores on the two-stage and pyramidal tests is high and primarily linear.

Table 5

Regression Analysis of Relationship between Two-stage and Pyramidal Scores (N=111)

Scoring Method	r	Regression of Two-stage Score on Pyramid Score		Regression of Pyramid Score on Two-stage Score	
		eta	p[a]	eta	p[a]
Number correct	.84	.85	.71	.88	.25
Mean difficulty--attempted	.81	.86	.10	.86	.04*
Mean difficulty--correct	.79	.84	.23	.84	.15
Difficulty of final item	.83	.83	.56	.86	.21

[a] Significance of curvilinearity
*Significant at p<.05

Internal Consistency Reliability

Table 6 shows the internal consistency reliabilities for the two-stage subtests. The internal consistency of the 10-item routing test (.72) was the same as that of the least difficulty 30-item measurement test. When number of items was equated for the 10-item routing test and the 30-item measurement tests, the routing test showed the highest internal consistency of the five subtests. This was likely due to the intentional restriction in the range of abilities of subjects assigned to each measurement test by the routing process.

Table 6

Internal Consistency Reliabilities for Subtests of the Two-stage Test

Subtest	N	Number of Items	Hoyt Reliability Coefficient
Routing	111	10	.72 (.89[a])
Measurement 1	21	30	.84
Measurement 2	20	30	.66
Measurement 3	27	30	.75
Measurement 4	43	30	.72

[a] Estimated reliability for a 30-item test.

While this finding might have resulted from differences in item discriminations among the subtests, comparison of the data in Table 1 with those in Table 6 show that the measurement tests with the highest average discriminations had the lowest reliabilities. Measurement test 1 (the most difficult measurement test) did, however, have a reliability which was almost as high as the corrected reliability of the routing test.

Mis-routing

In the two-stage test, only one testee in the sample of 111 obtained a score of 6 or less on the measurement subtest and was thus considered mis-routed. A less difficult measurement test would have been more appropriate for him/her. No perfect scores were obtained on any measurement subtest. The misclassification rate was therefore 1/111=.009.

Intercorrelations of Pyramidal Scores

The intercorrelations of scores from the four methods of scoring the pyramidal test are shown in Table 7. Highest observed correlation (r=.99) was between the two mean difficulty scores. Number correct had the lowest correlations (r=.93) with the two mean difficulty scores. There was no curvilinearity in these data since all the corresponding r's and etas were virtually identical.

Table 7

Intercorrelations of Scores from
Pyramidal Scoring Methods
(N=111)

Scoring Method	Number Correct	Mean difficulty--attempted	Mean difficulty--correct
Mean difficulty--attempted			
r	.93		
eta	.93		
Mean difficulty--correct			
r	.93	.99	
eta	.93	.99	
Difficulty of final item			
r	.98	.95	.95
eta	.98	.95	.96

DISCUSSION AND CONCLUSIONS

Score distributions for both the pyramidal and two-stage tests suggested that both were of appropriate difficulty for the general ability level of the testees. For the pyramidal tests, the mean score was slightly more than half of the possible range. Those pyramidal test scores which were expressed in standard units had means which were all about zero. The two-stage scores also had a near-zero mean. These results were similar to those obtained by Larkin and Weiss (1974) and Betz and Weiss (1973). However, the latter study found mean scores for a similar two-stage test to be slightly closer to zero (-0.21 at time 1 and -0.02 at time 2) than in the present study (-0.33). In the previous investigation of two-stage tests, standard deviations were found to be 1.36 and 1.39. In the present study, the standard deviation was 1.30.

A "real data" simulation of the same two-stage test used in the present investigation (Betz & Weiss, 1974) resulted in a mean score of very near zero (-.004) and a standard deviation of 1.05. Thus, real testees obtained a lower average score, and were more variable on the two-stage test, than were simulated testees. These results suggest that there are very few chance successes due to guessing in actual administration of two-stage tests, since guessing would result in scores above zero, on the average.

The two-stage test was found to utilize a smaller proportion of its possible score range (.13) than the pyramidal test. This finding is consistent with the results of the two previous empirical studies in this series, in which two-stage tests and two methods of scoring pyramidal tests used a greater proportion of the score range than conventional tests. Betz and Weiss (1973) found that, for a similar two-stage test, the proportion of range utilized was .23. However, their index was computed by dividing the obtained standard deviation by 6 (±3 s.d.) rather than the actual possible range of two-stage scores, thus inflating the index. The range of possible scores for the two-stage test in the present study was 9.96 rather than simply 6. Therefore, the proportion of range utilized is lower in the present study because of the change in the method of computation.

Both adaptive tests provided score distributions which were slightly skewed in a positive direction, but, with the exception of one scoring method for the pyramidal test, the degree of skew was not statistically significant. Seeley, Morton and Anderson (1962) obtained a highly negatively skewed distribution of scores on a pyramidal test. Their result, however, was possibly due to the easiness of their test and/or to the exclusion of some lower-ability examinees who did not carefully follow the instructions. Bayroff and Seeley's (1967) results, however, were more similar to those found in the present study; they obtained a normal distribution of pyramidal scores when computer administration was employed. Larkin and Weiss (1974) found a tendency toward positive skew in two other pyramidal tests similar to the one used here.

In their previous empirical study of two-stage testing, Betz and Weiss (1973) obtained score distributions which also tended toward positive skew but were not significantly different from a normal distribution. The two-stage simulation (Betz & Weiss, 1974) showed score distributions to have almost zero skew (-.04) when administered to a population distributed normally on ability.

There was a slight, but non-significant, trend for most pyramidal score distributions to be platykurtic. The tendency toward flatness in score distributions from pyramidal tests has been noted by Hansen (1969) who obtained a rectangular score distribution. Two similar pyramidal tests of Larkin and Weiss (1974) were significantly flat. The two-stage score distribution in the present study was slightly (and non-significantly) leptokurtic. Betz and Weiss (1973), however, found that a similar two-stage test produced a slightly flattened distribution of scores. With simulated data, Betz and Weiss (1974) found that score distributions on the same two-stage test used in the present study were significantly flat *(p<.01)*, but less platykurtic than distributions of scores for another two-stage test and a conventional test. Results of the present study, however, showed that the mean-difficulty-correct score derived from the pyramidal test gave results which were least deviant from a normal distribution in comparison to other pyramidal and two-stage test scores.

The distributions of scores within the two-stage measurement subtests represented an improvement over those obtained in the previous study of two-stage testing. First, the number of testees assigned to each measurement test was more nearly equal. Betz and Weiss (1973) found that approximately half of the subjects completing their two-stage test were routed to the most difficult measurement test. Further, the easier measurement tests in the previous study were found to be too easy for the testees routed to them. The more even distribution of testees routed to each measurement test in the present investigation can be attributed to the more appropriate difficulty of the routing test and to the revised procedure used to determine cutting scores for assignment to measurement tests. The improvement in the score distributions within the measurement tests is due to modifications making the more difficult measurement tests easier, and the less difficult measurement tests more difficult.

The misclassification rate for the two-stage test in this study was .009 using the same criteria as those used by Betz and Weiss (1973), i.e., perfect scores (30) or chance scores (or less) on the measurement tests. This compared favorably with the 5% misclassification rate in Betz and Weiss (1973). The 20% rates obtained by Angoff and Huddleston (1958) and by Cleary, et al. (1969a,b; Linn, et al., 1969) were due primarily to the different misclassification criteria in their real-data simulation studies. The low rate of misclassifications in the present study may be accounted for by (1) the more accurate assignment of subjects to measurement tests brought about by revisions in the routing tests, (2) the maximum likelihood procedure used for classification, (3) the increased cutting scores (no testee was routed to a measurement test in which he/she obtained a perfect score) and (4) the more appropriate difficulties of the items used in the measurement tests.

The internal consistency reliabilities of the two-stage subtests also reflect the improvements in the difficulties of those subtests. For the routing test and three of the four measurement tests, measures of internal consistency were as much as .31 higher than the corresponding reliabilities found by Betz and Weiss (1973). This finding suggests that the difficulties of the measurement test items were more appropriate (i.e., approximating p=.5) for the groups of subjects attempting them. The increased difficulty of the routing test items in the present study as compared to the previous empirical study resulted in routing test scores which had a standard deviation more than twice that found in the previous study. The changes made in the measurement tests, by decreasing the number of items which were much too easy or too difficult for the group routed to them, enabled the interitem correlations and thus the internal consistency reliability coefficient to increase.

The correlations between scores on the pyramidal and two-stage tests obtained in this study ranged from r=.79 to .84 (eta=.83 to .88). The two previous empirical studies in this series found correlations of r=.82 to .89 (eta=.84 to .92) between scores on the pyramidal and conventional testing strategies, and r=.80 to .84 (eta=.82 to .88) between scores on the two-stage and conventional tests. In the simulation study, Betz and Weiss (1974) found a correlation of r=.82 (eta=.82) between scores on the two-stage and conventional tests. Thus, it appears that scores on the two-stage test are almost as highly related to scores on a 15-stage pyramidal test as they are to scores on a 40-item conventional test. The relationship between scores on the two adaptive tests is almost as high as that between the pyramidal and conventional tests. In the two previous empirical studies, the items contained in the adaptive and conventional tests were non-overlapping. The present study, however, permitted some of the same items to be administered in both the pyramidal and two-stage tests, which may have somewhat inflated the correlation between them. An average of six items--or 40% of the pyramidal test's items--were the same in both tests.

The correlation between scores on the two adaptive strategies approached their empirical stabilities. The seven-week test-retest stability of the pyramidal test used in this study ranged from r=.82 to .86 (eta=.85 to .90) depending on the scoring method used (Larkin & Weiss, 1974). The stability of scores of a two-stage test similar to the one used here was r=.88 (Betz & Weiss, 1973). Using the Pearson coefficients, the correlation between the two-stage and pyramidal tests accounted for 62% to 71% of the common variance. Stability of the adaptive tests showed that from 67% to 74% of the pyramidal test's variance was reliable while about 77% of the variance of the two-stage test was reliable. Thus, assuming that error variance is uncorrelated, from 42% to 53% of the reliable variance in the pyramidal test was common to the two-stage test, while from 48% to 55% of the reliable variance in the two-stage test was common to the pyramidal test. Further, the correlation between the two adaptive tests equalled or exceeded the internal consistency reliabilities of all the measurement tests and approached the internal consistency of the routing test when corrected for length.

Several tenative conclusions can be drawn from these results. First, the results replicate previous findings which indicate that the order of administration of adaptive tests does not significantly affect scores on the tests. Consequently, research on different adaptive strategies can proceed by administering two or more strategies successively to an individual without randomizing administration order.

The results seem to support previous findings by Lord (1970) and Larkin and Weiss (1974) which indicate that the average difficulty scores are the most useful way of scoring pyramidal tests. Lord's results indicate that his average difficulty score provides the most desirable information functions while Larkin and Weiss' results indicate that these scores are the most stable over short time intervals. And, in the present study, the mean-difficulty-correct score gave results which deviated least from a normal distribution.

Although the distribution of ability in the subjects was unknown, this agreement of results across these studies implies that it is not unreasonable to assume that it was normal. Further research is needed, however, with populations of known distribution of ability, to support this assumption.

The data on score means for the two adaptive strategies suggest that few chance successes occurred, on the average, as the result of guessing. These results support Hansen's (1969) finding that decreases in guessing do occur when item difficulties are adapted to each individual's ability level. There was a suggestion in the data that the pyramidal strategy appeared to result in fewer chance successes due to guessing than did the two-stage strategy. This finding should also be further studied by research designed specifically to answer that question.

Finally, the results suggest that the two adaptive strategies are not replacements for each other in terms of measuring the same variable in the same way. When the correlation between scores on the two adaptive strategies was considered with respect to available data on the reliabilities of the strategies, only about 50% of the reliable variance of the two strategies was found to be common. Thus, each strategy orders individuals differently on estimated ability. Further research is needed to determine the reasons for these different ability estimates.

Thus, a deficiency of the present study concerns the determination of the relative efficiency of the two testing strategies. The use of live subjects does not permit any estimation of the precision or accuracy of the scores obtained under either strategy, since the "true" ability of the testees was, of course, unknown. Thus, the degree to which test scores accurately reflected underlying ability could not be determined. Live-testing empirical studies designed to answer this question will require very large samples of testees. Theoretical studies, as shown by Weiss and Betz (1974), appear to provide results which are not generalizeable beyond those conditions satisfying their restrictive assumptions. Thus, additional simulation studies (e.g., Betz & Weiss, 1974) seem to be necessary to determine which adaptive tests scored by which method provide most accurate measurement for testees of various ability levels. The simulation studies should then be followed by live-testing studies to validate the simulation findings.

References

Angoff, W.H. & Huddleston, E.M. The multi-level experiment: a study of a two-level test system for the College Board Scholastic Aptitude Test. Princeton, New Jersey: Educational Testing Service, Statistical Report SR-58-21, 1958.

Bayroff, A.G. Psychometric problems with branching tests. Paper presented at the meeting of the American Psychological Association, Division 5, September, 1969.

Bayroff, A.G. & Seeley, L.C. An exploratory study of branching tests. U.S. Army Behavioral Science Research Laboratory, Technical Research Note 188, June, 1967.

Bayroff, A.G., Thomas, J.J. & Anderson, A.A. Construction of an experimental sequential item test. Research memorandum 60-1, Personnel Research Branch, Department of the Army, January, 1960.

Betz, N.E. & Weiss, D.J. An empirical study of computer-administered two-stage ability testing. Research Report 73-4, Psychometric Methods Program, Department of Psychology, University of Minnesota, October, 1973. (AD 768993)

Betz, N.E. & Weiss, D.J. Simulation studies of two-stage ability testing. Research Report 74-4, Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, 1974. (AD A001230)

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)

DeWitt, L.J. & Weiss, D.J. A computer software system for adaptive ability measurement. Research Report 74-1, Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, 1974. (AD 773961)

Hansen, D.N. An investigation of computer-based science testing. In R.C. Atkinson and H.A. Wilson (eds.), Computer-assisted instruction: a book of readings. New York: Academic Press, 1969.

Hoyt, C.J. Test reliability estimated by analysis of variance. Psychometrika, 1941, 3, 153-160.

Larkin, K.C. & Weiss, D.J. An empirical investigation of computer-administered pyramidal ability testing. Research Report 74-3, Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, 1974. (AD 783553)

Linn, R.L.,, Rock, D.A. & Cleary, T.A. The development and evaluation of several programmed testing methods. Educational and Psychological Measurement, 1969, 29, 129-146.

Lord, F.M. Some test theory for tailored testing, In W.H. Holtzman (ed.), Computer-assisted instruction, testing, and guidance. New York: Harper and Row, 1970.

Lord, F.M. Robins-Monro procedures for tailored testing. Educational and Psychological Measurement, 1971, 31, 3-31. (a)

Lord, F.M. Tailored testing, an application of stochastic approximation. Journal of the American Statistical Association, 1971, 66, 707-711. (b)

Lord, F.M. A theoretical study of Two-stage testing. Psychometrika, 1971, 36, 227-241. (c)

Lord, F.M. & Novick, M.R. Statistical theories of mental test scores. Reading, Mass.: Addition-Wesley, 1968.

McBride, J.R. & Weiss, D.J. A word knowledge item pool for adaptive ability measurement. Research Report 74-2, Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, 1974. (AD 781894)

McNemar, Q. Psychological statistics (4th ed.). New York: Wiley, 1969.

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

Seeley, L.C., Morton, M.A. & Anderson, A.A. Exploratory study of a sequential item test. U.S. Army Personnel Research Office, Technical Research Note 129, 1962.

Waters, C.W. Preliminary evaluation of simulated branching tests. U.S. Army Personnel Research Office, Technical Research Note 140, 1964.

Waters, C.W. & Bayroff, A.G. A comparison of computer-simulated conventional and branching tests. Educational and Psychological Measurement, 1971, 31, 125-136.

Weiss, D.J. The stratified adaptive computerized ability test. Research Report 73-3, Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, 1973. (AD 768376)

Weiss, D.J. Strategies of Adaptive ability measurement. Research Report 74-5, Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, 1974.

Weiss, D.J. & Betz, N.E. Ability measurement: conventional or adaptive? Research Report 73-1, Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, 1973. (AD 757788)

Appendix A

Difficulty (b) and Discrimination (a) Item Parameters for the Two-stage Test

Routing Test			Measurement Test 1			Measurement Test 2			Measurement Test 3			Measurement Test 4		
Item Reference Number	b	a	Item Reference Number	b	a	Item Reference Number	b	a	Item Reference Number	b	a	Item Reference Number	b	a
161	-.26	.86	328	2.31	.54	306	.97	.49	145	-.41	.59	204	-1.15	.73
661	-.30	.58	166	2.03	.64	321	.79	.63	292	-.58	.48	94	-1.57	.49
670	-.28	.62	309	2.47	.48	660	1.01	.40	382	-.48	.64	642	-1.80	.42
52	-.28	.61	298	2.62	.43	114	.65	.77	205	-.62	.47	109	-1.06	.89
599	-.23	.81	627	2.67	.42	630	-.05	1.31	207	-.536	.60	515	-1.33	.62
329	-.21	.86	662	1.93	.57	231	.79	.45	137	-.74	.40	141	-1.83	.42
144	-.18	.63	385	2.35	.42	656	.71	.44	46	-.81	.67	108	-1.71	.47
50	-.23	.50	336	2.04	.49	215	.65	.48	203	-.84	.65	87	-1.10	.99
369	-.22	.56	297	2.31	.40	651	.49	.56	33	-.85	.64	276	-2.12	.41
272	-.13	.98	274	2.13	.42	296	.34	.91	53	-1.01	.52	43	-1.21	.90
			180	2.07	.43	666	.42	.55	188	-.47	.71	214	-2.08	.42
Mean	-.23	.70	245	2.32	.38	375	.46	.49	365	-.56	.66	640	-1.47	.67
S. D.	.05	.16	381	1.79	.50	111	.46	.48	234	-.69	.51	285	-1.42	.71
			273	1.79	.49	340	.30	.78	154	-.58	.66	36	-1.08	1.23
			319	1.49	.62	302	.37	.50	208	-.68	.58	637	-1.40	.75
			359	1.54	.58	271	.33	.53	156	-.63	.65	47	-1.31	.87
			115	1.88	.45	264	.21	.86	270	-.52	.86	232	-1.70	.59
			360	2.18	.34	60	.24	.66	143	-.57	.77	173	-1.43	.76
			652	1.33	.60	113	.25	.61	667	-.73	.57	641	-1.89	.52
			152	1.40	.55	283	.15	.97	211	-.72	.61	189	-1.60	.66
			378	1.44	.49	265	.17	.77	224	-.79	.54	649	-2.21	.44
			263	1.38	.51	386	.14	.70	91	-.59	.83	103	-1.34	.89
			120	1.07	.72	146	.00	.61	37	-.69	.67	88	-1.75	.63
			174	1.16	.64	633	-.08	.50	390	-.73	.63	227	-1.63	.71
			140	1.30	.52	568	-.08	.91	221	-.74	.65	86	-1.55	.77
			288	1.11	.56	59	.17	.64	307	-.84	.56	40	-1.34	1.02
			162	1.17	.52	315	.17	.83	58	-.96	.48	199	-1.42	.92
			337	.73	.98	342	.17	.77	588	-.89	.53	95	-2.20	.50
			294	.79	.70	266	.16	.86	155	-1.35	.34	311	-1.83	.66
			299	.98	.52	347	.14	1.07	535	-.68	.86	643	-2.56	.44
			Mean	1.72	.53	Mean	.35	.68	Mean	-.71	.61	Mean	-1.60	.68
			S.D.	.56	.13	S.D.	.30	.21	S.D.	.19	.12	S.D.	.37	.21

Appendix B

Difficulty (b) and Discrimination (a) Item Parameters for the Pyramidal Test

Stage																													
1															b= -.05 a= *1.31*														
2														-.21 *.86*		.14 *1.07*													
3													-.52 *.86*		-.13 *.98*		.34 *.91*												
4												-.70 *1.82*		-.25 *.86*		.15 *.97*		.49 *.56*											
5											-.85 *.75*		-.59 *.83*		-.08 *.91*		.21 *.86*		.73 *.98*										
6										-1.08 *1.23*		-.73 *.92*		-.23 *.81*		.16 *.86*		.42 *.55*		.98 *.52*									
7									-1.23 *1.35*		-.91 *.67*		-.57 *.77*		-.18 *.63*		.30 *.78*		.65 *.77*		1.07 *.72*								
8								-1.51 *1.40*		-1.10 *.99*		-.68 *.86*		-.29 *.75*		.17 *.83*		.46 *.49*		.97 *.49*		1.33 *.60*							
9							-1.68 *1.46*		-1.33 *1.16*		-.81 *.67*		-.47 *.71*		-.08 *.50*		.24 *.66*		.79 *.70*		1.16 *.64*		1.49 *.62*						
10						-1.99 *1.76*		-1.42 *.92*		-1.06 *.89*		-.75 *.82*		-.40 *.68*		.17 *.77*		.46 *.48*		.98 *.38*		1.30 *.52*		1.79 *.51*					
11					-2.13 *1.10*		-1.67 *1.02*		-1.34 *1.02*		-.84 *.65*		-.56 *.66*		-.04 *.47*		.25 *.61*		.79 *.63*		1.11 *.56*		1.54 *.58*		1.89 *.85*				
12				-2.22 *1.52*		-1.87 *1.43*		-1.55 *.77*		-1.10 *.77*		-.74 *.65*		-.28 *.62*		.17 *.77*		.47 *.26*		.83 *.37*		1.38 *.51*		1.79 *.49*		2.03 *.64*			
13			-2.41 *3.00*		-2.19 *.56*		-1.71 *.99*		-1.34 *.89*		-.85 *.64*		-.48 *.64*		-.17 *.47*		.33 *.53*		.65 *.48*		1.17 *.52*		1.40 *.55*		1.93 *.57*		2.31 *.54*		
14		-2.72 *3.00*		-2.22 *1.07*		-1.92 *1.23*		-1.43 *.76*		-1.07 *.76*		-.63 *.65*		-.28 *.61*		.07 *.76*		.48 *.22*		.92 *.37*		1.31 *.44*		1.65 *.39*		2.05 *.49*		2.47 *.47*	
15	-2.86 *1.01*		-2.41 *3.00*		-2.20 *.51*		-1.66 *.93*		-1.31 *.87*		-.89 *.53*		-.53 *.60*		-.09 *.41*		.37 *.51*		.79 *.45*		1.01 *.42*		1.44 *.49*		1.88 *.45*		2.35 *.42*		2.61 *.43*
Difficulty Level	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29