AN EMPIRICAL INVESTIGATION OF COMPUTER-ADMINISTERED PYRAMIDAL ABILITY TESTING

Kevin C. Larkin

and

David J. Weiss

Research Report 74-3

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

July 1974

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

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Research Report 74-3	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): An Empirical Investigation of Computer-administered Pyramidal 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, VA 22217	12. REPORT DATE: July 1974
	13. NUMBER OF PAGES: 59
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified
	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	sequential testing	programmed testing
ability testing	branched testing	response-contingent testing
computerized testing	individualized testing	automated testing
adaptive testing	tailored testing	

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

Three pyramidal adaptive tests and a conventional peaked test were constructed and administered by time-shared computer to two separate groups of students enrolled in undergraduate psychology courses. Six different methods of scoring pyramidal tests were evaluated with respect to score distributions, stability, and the degree of relationship among scoring methods and between pyramidal scoring methods and scores on the (over)

20 (continued)

conventional test. For both the pyramidal tests and the conventional test, score distributions were platykurtic and positively skewed. Two methods of scoring the pyramidal tests consistently used an equal or greater proportion of the range of possible scores than the conventional test. The 15-stage pyramidal tests showed test-retest correlations which were only slightly lower than that for the 40-item conventional test. However, when the effects of memory were considered, the pyramidal strategy yielded more stable ability estimates than conventional tests of equivalent length. The correlation between pyramidal test scores and those on conventional tests ranged from .82 to .86 depending on the scoring method used. One pair of scoring methods was found to be perfectly correlated for properly constructed pyramidal tests; a second pair correlated almost perfectly. Findings were generally in favor of pyramidal testing, but further investigation of this adaptive testing strategy seems necessary to determine its other important psychometric characteristics and to develop optimal rules for constructing pyramidal item structures.

Contents

Contents, continued

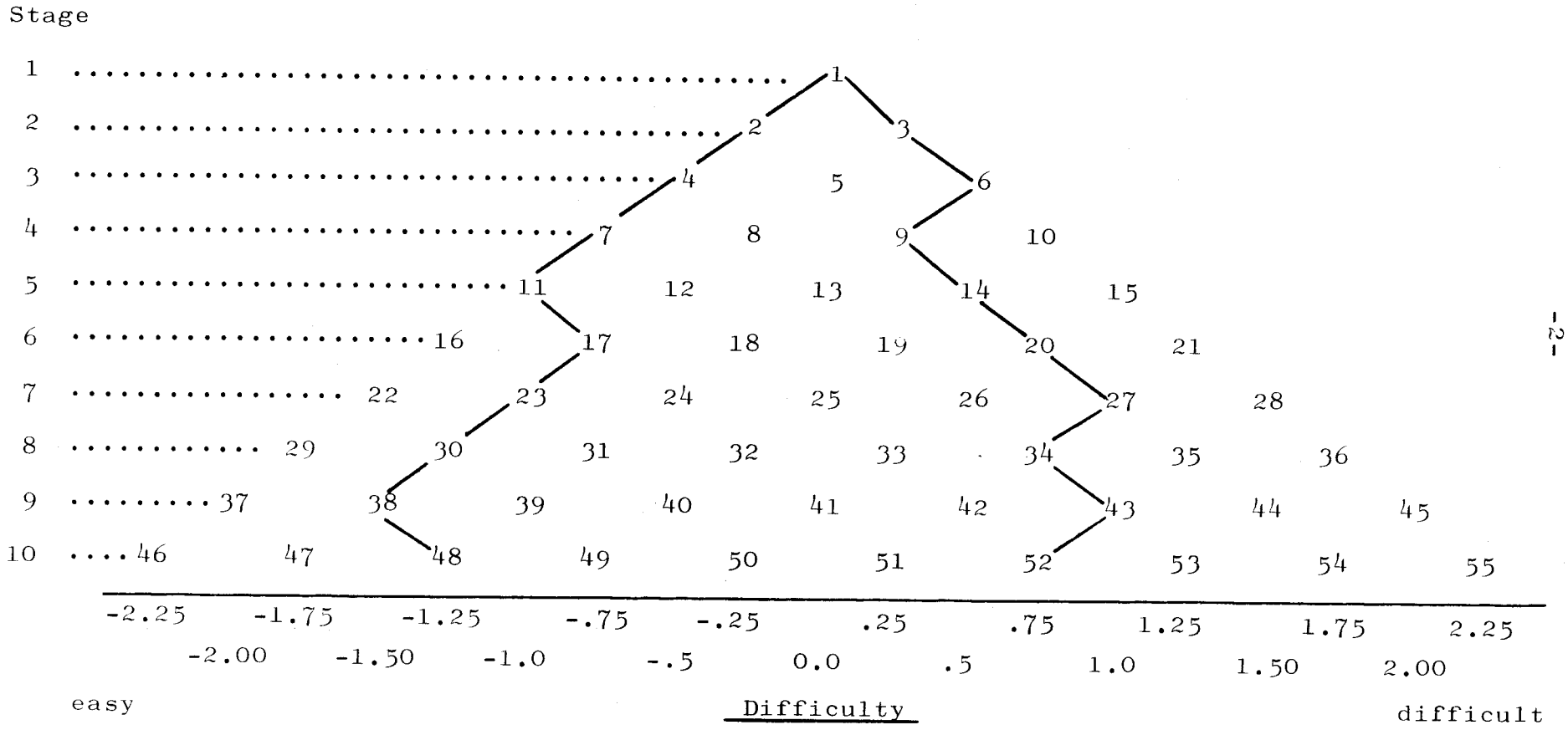
AN EMPIRICAL INVESTIGATION OF COMPUTER-ADMINISTERED PYRAMIDAL ABILITY TESTING

Conventional tests of ability have traditionally been administered by paper and pencil to large groups of individuals. Each subject is expected to attempt every item in the test regardless of its difficulty or his/her ability. Administration of ability test items by interactive computer systems has made possible the tailoring of tests to the ability of the individual testee. When an ability test is administered by computer, items are selected for presentation according to a pre-determined set of rules or "strategy" which takes into account the testee's responses to previously administered items. Adaptive testing strategies are differentiated by the set of rules used to determine item selection (Weiss, 1974). The rationale for adaptive testing is that, by eliminating those items which are either too difficult or too easy for the person taking a test, its reliability and validity may be improved and testing time shortened. Weiss and Betz (1973) have described the various strategies used and have summarized the research literature on adaptive testing.

The strategy most frequently used in adaptive testing has been called "branched", "sequential", or "pyramidal" testing. This method requires that items be arranged in a triangular structure according to difficulty. Figure 1 illustrates a pyramidal item structure. Typically, the first item administered (item 1, stage 1) is of median difficulty for the group taking the test, and is represented at the top of the pyramidal structure. The second item presented (stage 2) is contingent upon whether the response to the first item was correct or incorrect. If the testee answers the first item correctly, an item of greater difficulty (item 3) is administered next. An incorrect response to item 1 results in the administration of a second-stage item of lesser difficulty (item 2). Thus, as Figure 1 shows, there are two items at the second level or "stage" of the pyramid. The testee is routed to an item at stage 3 according to his response to the stage 2 item; again a more difficult item follows a correct response, and an easier item follows an incorrect response. The branching procedure is repeated until the subject has attempted one item at each of a fixed number of stages. The solid lines connecting item numbers in Figure 1 illustrates the paths of two hypothetical testees through the pyramidal structure.

The number of items attempted by a testee is equal to the number of stages (provided that one item is administered at each stage), and is only a fraction of the total number of items needed to construct the pyramidal structure. In

Figure 1. A ten-stage pyramidal adaptive test structure.



the pyramidal structure shown in Figure 1, each testee would encounter only 10 of the 55 items available for administration. Many variations of this method of testing have been suggested (Weiss, 1974). For example, the number of items to be administered at one stage may be set at three or five. In such cases, branching is based on the number of items answered correctly at a given stage. Instead of routing from item to item, the testee is branched from one block of items to another with all items in a block having about the same difficulty.

The increment or decrement in the difficulty of items at one stage to those in the next (.25 in Figure 1) is called the "step size" and may be either fixed or variable. Some pyramidal tests (Paterson, 1962; Lord, 1971a) have used a large step size at the beginning of the test to make relatively coarse distinctions among ability levels; as testing proceeds the step size becomes smaller or "shrinks" enabling finer and finer discriminations among testees. In most cases, the increment in difficulty for a correct response is equal to the decrement in difficulty following an incorrect response. This insures symmetric branching throughout testing, and requires that one item at each stage be attempted. This has been called an "up-one /stage/ /down-one /stage/" strategy, or "equal offset".

The term "unequal offset" has been used to explain branching which is asymmetric (Lord, 1970). In such a case, following a correct response a testee is routed to a more difficult item in the next stage, but after an incorrect response, routing occurs to a much easier item two or even three stages further into the pyramid (i.e., one or more stages is skipped). This is known as an "up-one/down-two (or -three)" strategy and is most commonly used as a correction for guessing. In this variation the number of items administered is less than the number of stages, unless the testee responds correctly to all items administered.

Pyramidal tests may be scored by a number of different methods. First, the rank of the difficulty of the final item attempted can be considered the individual's score (Bayroff, Thomas & Anderson, 1960; Seeley, Morton & Anderson, 1962; Waters & Bayroff, 1971). The pyramidal test illustrated in Figure 1 would, therefore, yield 10 scores. The number of ranks may be doubled by assigning a higher rank to a subject answering the final item correctly, than to one who does not (Waters, 1964; Bayroff & Seeley, 1967). The difficulty level of the final item reached (e.g., Bayroff, 1969) may also be considered an estimate of a testee's ability (e.g., -1.5 and +1.0 for the two testees shown in

Figure 1). Another method, which takes into account the correctness or incorrectness of the response to the final item involves branching the subject to an hypothetical item following the last item administered and estimating its difficulty. This has been named the "final node score" (Hansen, 1969) or "final difficulty score" (Lord, 1971b). To distinguish this method from the one utilizing the difficulty of the last item, it can be called the "n + 1^{th} item" scoring method. Another scoring method involves the average of all items attempted or all items correctly answered. Lord (1970) has used a related averaging method which eliminates the first item (since everyone attempts it) but includes the n + 1^{th} item. He considers it the "score of choice" (Lord, 1971b, p. 709) for most up-one/down-one strategies. Finally, a more complicated scoring system has been proposed by Hansen (1969) which assigns an estimated score to each item in the pyramid.

Empirical studies. Early research with pyramidal tests used paper and pencil administration. Krathwohl and Huyser (1956) administered an eight-stage (one item per stage) and a four-stage (two items per stage) pyramid to 100 college students. They obtained correlations of .78 and .68 between the pyramidal tests and 60-item parent tests. Their pyramidal tests were completed more quickly than the conventional tests, and provided almost as much information.

Bayroff, Thomas and Anderson (1960), following Krathwohl's approach, constructed four six-stage pyramidal tests using a decreasing step size. Based on their response choice on the first item testees were routed to one of three alternative items at stage 2. Those who selected the correct alternative were administered a more difficult item; those who responded with either of two plausible distractors were routed to an item of the same difficulty as the initial item; and those who chose the least popular incorrect response were given an easier item. For the remaining stages, ordinary up-one/down-one branching was used. Seeley, Morton and Anderson (1962) administered these six-stage pyramidal tests to 327 men and correlated scores on the pyramidal tests with those obtained on corresponding subtests of a longer conventional test. For both verbal and numeric items, the correlation between the pyramidal and conventional tests was .63; however, the distribution of pyramidal scores was highly skewed with a large number of scores at the high end of the distribution. These authors also reported that a number of the low ability testees did not follow the routing instructions, resulting in unusable test records for these examinees.

Wood (1969) administered paper and pencil pyramidal tests of 4, 5, and 6 stages to 91 students. Step size was fixed at p = .05; the initial item was of median difficulty (p = .50); an up-one/down-one branching rule was used; and the score was the number of items correctly answered in each test. Validity of the tests was determined by correlations of test scores with course grades in comparison with those obtained with a 46-item conventional test. Correlations between the pyramidal scores and course grades were all below .35; combining scores on the three pyramidal tests increased the correlation to .51. The correlation between the conventional test and grades was .68, and a test composed of the fifteen most discriminating items in the conventional test had a correlation of .52 with course grades. Wood concluded that a conventional test is just as good as a combination of pyramidal tests composed of the same number of items.

More recent empirical studies have used computers to administer adaptive tests. Bayroff and Seeley (1967) administered two eight-stage pyramidal tests by teletype to 102 men. The step size used was p = .05 and final item difficulties ranged from p = .95 to p = .20; scores were based on the correctness or incorrectness of the final item, providing a score range of 17 points. Testees also completed 40-item numerical and 50-item verbal conventional tests. Correlations between the adaptive and conventional tests were .83 and .79 (corrected for restriction of range) compared to an estimated correlation between eight-item conventional tests and the 40- and 50-item conventional tests of .75 and .67. Thus, pyramidal tests proved to be more highly related to the long conventional tests than were conventional tests of comparable length. By use of the Spearman-Brown formula, it was found that conventional tests would require at least twice as many items as the pyramidal tests to achieve the same correlation with the criterion paper and pencil tests.

Hansen (1969) administered five different pyramidal tests by teletype to 56 college freshmen. The number of stages per test was either three or four with each student answering a total of 17 items. Hansen used a step size of p = .10 and scored his tests by four different methods. Scores on the pyramidal tests were correlated with scores on a one-hour classroom exam on the same material completed one week before the pyramidal tests were administered, and with scores on another achievement test and final course grade. The conventional test, even when equated for length, was found to have a lower internal consistency reliability than any of the five pyramidal tests. Scores for the pyramidal

tests were distributed more rectangularly than those of the conventional test which had a negatively skewed distribution. Results also showed that the pyramidal tests were completed in an average of five minutes less time than the conventional test. Pyramidal tests scored by two methods also showed higher correlations than the conventional test with final grade and the achievement test criterion. A second study produced similar results.

Bryson (1971) compared two five-stage pyramidal tests with two five-item conventional tests on their correlation with 100-item parent tests. Conventional tests were administered by paper and pencil while the pyramidal tests were administered using a cathode ray computer terminal. In one of the pyramidal tests, the item selection procedure sequentially selected items based on the most discriminating item for all those who reach a given point in the pyramidal structure, while the other used an item selection procedure designed to maximize the prediction of total score (Wolfe, 1970). Both pyramids had a variable step size. Each pyramidal strategy was administered to two groups of 263 subjects and the conventional tests were administered to comparable groups of 250 individuals. Results indicated that one of the short conventional tests was more highly correlated with total test score than either of the pyramidal tests. One of the pyramids had lower correlations with total test score than either of the conventional tests.

Simulation studies. Simulation involves scoring a conventional test "as if" it had been administered adaptively (real data simulation) or using computers to generate hypothetical subjects, items, and/or test response records (computer simulation). Bryson's (1971) investigation compared her empirical findings with those of a real data simulation using the same four pyramidal and conventional tests with two groups of 100 subjects. The highest correlations with total test score were obtained when one of the two pyramids was used. The other pyramidal strategy had correlations less than or equal to one of the conventional tests and higher correlations than the other. These findings were more favorable to adaptive testing than her empirical results.

Linn, Rock and Cleary (1969) investigated seven different branching strategies using real data simulation based on the responses of 4,885 students to a 190-item conventional test. For each strategy, the appropriate items from the longer tests were selected and scored as if the testees had attempted only those items in the order required by the given adaptive test. Five of the simulated branching strategies

were two-stage procedures (Betz & Weiss, 1973); the two remaining designs were pyramidal. The first was a ten-stage pyramid with a step size of about p = .02. The second pyramidal test consisted of five stages with five items per stage; thus, 25 items were attempted by each subject with branching based on a subject's performance within each block. Both pyramids used an equal offset. Pyramidal tests were compared to five shortened conventional tests of from 10 to 50 items. Results showed that the 10-stage pyramidal test correlated .87 with total test score; the 25-item pyramid correlated .95; and the short conventional tests correlated .89 to .96. The 25-item pyramid's correlation with total test score corresponded to that of a 35-item conventional test. Linn et al. (1969) also obtained scores on two achievement tests for the same subjects, which were used as criterion measures. The 10-item pyramidal test showed a higher correlation with the criterion measures than the conventional test of the same length. Similarly, the five-stage 25-item pyramid correlated higher with the criterion tests than the 50-item conventional test. These findings imply that pyramidal testing can result in gains in validity with fewer items administered in comparison to conventional testing.

Paterson (1962) conducted a monte carlo computer simulation study using a pyramidal strategy. Items in the pyramid were first structured by difficulty and then ordered by discriminations. The first items administered were the most discriminating while the later items were less discriminating within each level of difficulty. Step size varied as a function of item discrimination. If a highly discriminating item was answered correctly, the increment in difficulty between that item and the next was large. When an item of low discrimination was answered correctly, the increment in difficulty was small. Similarly, decrements in step sizes depended on the discriminations of items which were answered incorrectly. Since items were arranged according to discriminations, the step sizes at the beginning of the test were large and decreased as the testee moved through the pyramidal structure.

Paterson's pyramid consisted of six stages and was compared with a six-item conventional test for an hypothetical population of 1,500 individuals, with 100 people at each of 15 ability levels. The two testing strategies were compared at five levels of item discrimination under conditions of normal, rectangular, and U-shaped distributions of ability. The effects of errors in estimating the item parameters were studied by including items of inappropriate difficulty or discrimination in the pyramidal tests. The data led to the conclusion that errors in parameter estimates in pyramidal testing did not seriously affect the score distributions

obtained. Pyramidal testing was found to give better estimates of ability than conventional tests when U-shaped or rectangular distribution of ability were assumed. Pyramidal test scores were also more precise than conventional scores, especially at the extremes of the ability distribution, and could predict ability from test scores as well as conventional tests.

Theoretical studies. Waters (1964) conducted a theoretical comparison of a five-stage pyramidal test and four conventional five-item tests using Lord's (1952) model to obtain the correlation between test score and underlying ability for each test. The hypothetical pyramidal test used a step size of p = .10, an up-one/down-one branching rule, and was scored by two methods. Under either scoring method, the correlation between test score and ability was higher for the pyramidal test than for any of the conventional tests, whether free-response or multiple-choice format was used. The pyramidal test produced a more rectangular score distribution and a potentially greater dispersion of scores than the conventional tests.

Waters and Bayroff (1971; Waters, 1970) compared 5-, 10-, and 15-stage pyramids and a ten-stage pyramid with two items per stage to conventional tests of the same length. Both conventional tests and pyramidal tests differed in the variability of item difficulties, and item discriminations were systematically varied. The distribution of ability was assumed to be normal. Results showed correlations of test score and ability were related to both the distribution of item difficulties and item discrimination, that correlations for the pyramidal tests were higher than those for the conventional tests, particularly with highly discriminating items, and that the one-item-per-stage pyramids showed higher correlations of test scores and ability than the two-item-per-stage pyramids.

Lord has reported several theoretical studies on pyramidal testing (Weiss & Betz, 1973). His analyses, based on the mathematics of item characteristic curve theory and the theory of Markov chains, compared 10-, 15-, and 60-stage pyramids with conventional tests of 60 items (Lord, 1970. 1971a, b; Stocking, 1969). Step sizes were systematically varied across tests but remained constant for any given test. Branching rules studied were up-one/down-one, up-one/down-two, up-one/down-three, and up-two/down-three, under a variety of scoring methods. Results showed that for conventional tests the information function was bell-shaped, leptokurtic, and symmetric about the median ability level; ability was most accurately estimated from test scores for those subjects at or near the median ability. Pyramidal information functions

were platykurtic, in some cases approximating a straight line, indicating that precision of test scores was more nearly equal across ability levels. At the median ability level, the 60-item conventional test provided more precise measurement than any pyramidal test. However, for abilities beyond $\pm.5$ to ±1.0 standard deviations the pyramidal tests provided more precise measurement. Different methods of scoring the pyramid provided different results, as did different stepping rules. Lord (1970, 1971a) also investigated a variable step size procedure adapted from bio-assay work called the Robbins-Munro procedure. In this strategy large increments or decrements in item difficulty occur early in the testing process with progressively smaller step sizes occurring later in testing. The procedure is designed to converge on a difficulty level at which each individual has a .50 probability of answering each item correctly. Although this procedure yielded extremely favorable results for pyramidal tests, it requires item pools that are so large as to be practically unfeasible.

Mussio (1972) has attempted to reduce the large number of items required in pyramidal testing by adopting "reflecting barrier" and "retaining barrier" strategies. Both modifications involve truncating the upper and lower tails of the pyramidal structure, thus eliminating many items at extreme difficulty levels. Like Lord, Mussio presented his theoretical results in the form of information curves and obtained similar results. Pyramidal tests modified by either "barrier" provide less information at the mean of an ability distribution than a conventional peaked test, but much more information for those individuals whose ability deviates from the mean. The retaining barrier was found to provide more nearly equal estimates of precision over the range of abilities than the reflecting barrier. Although both approaches showed some loss in precision at very extreme ability levels, each was still more precise than conventional tests at those ability levels.

Summary. The research available on pyramidal testing has used a wide variety of subjects, item pools, and test characteristics including variations in branching strategies, entry points, step sizes, offsets, and scoring methods. Administration of considerably fewer items has resulted in shorter testing times when complex instructions and paper and pencil formats have been eliminated. Several pyramidal tests have shown higher correlations with parent tests than conventional tests of the same length. Pyramidal tests designed by Hansen (1969) and Linn et al. (1969) have obtained higher correlations with outside criteria than conventional tests. Pyramidal tests have also been shown to produce a more rectangular equidiscriminating score distribution than conventional tests (Hansen, 1969), and have higher correlations with underlying ability (Waters & Bayroff, 1971) when

the items are highly discriminating. Theoretical studies have also shown that pyramidal tests have nearly constant precision of measurement across all levels of ability. This level of precision is much greater than that for conventional tests at the more extreme ability levels (Lord, 1970, 1971a, b; Mussio, 1972; Paterson, 1962).

Much of the empirical and simulation research has attempted to determine how highly pyramidal tests correlate with longer conventional parent tests. Investigators have been concerned with constructing short adaptive tests which yield essentially the same information as a conventional test. The theoretical studies have demonstrated that, for many people, pyramidal tests may be more accurate measurement instruments than conventional tests. If this is the case, then the demonstration of a strong relationship between the two testing strategies is not of primary importance. One major purpose of adaptive testing is to obtain measures of ability which are more precise than those of conventional tests. When this is considered, a high adaptive-conventional correlation is neither necessary nor desirable.

None of the studies to date has attempted to assess the relative test-retest stabilities of pyramidal and conventional tests. Furthermore, only Hansen (1969) has studied the relationships between the various pyramidal scoring methods. The present investigation was designed to supplement the existing literature on pyramidal tests in these areas, and to replicate some of the findings of earlier studies using longer pyramids than had been used in previous empirical studies.

Method

The pyramidal tests used in this study represent only one of several strategies of adaptive testing being used in a larger series of research studies (e.g., Betz & Weiss, 1973). This series of studies is designed to investigate the possible advantages of adaptive testing strategies as compared to conventional ability testing procedures, and to determine which adaptive approaches provide the most accurate measurement of ability. Adaptive tests are being compared to conventional tests and to other adaptive strategies with respect to ability estimation, stability, internal consistency reliabilities, and other psychometric characteristics. At the same time, the research is concerned with answering basic questions about each adaptive strategy. These include such questions as optimum ways of structuring the branching paradigm, problems in determining branching rules, and determination of useful and reliable methods of scoring the adaptive tests.

All adaptive and conventional tests were administered by computer (DeWitt & Weiss, 1974). Testing strategies were administered two at a time so that scores on one adaptive test could be compared with those on another, and so that adaptive and conventional tests could be compared. Each individual was tested on two occasions with a period of about seven weeks between the initial and final testings, in order to compare the test-retest stabilities of each testing strategy, and scoring methods within a strategy.

Test Development

Item Pool. The item pool consisted of 369 five-alternative multiple-choice vocabulary questions (see McBride & Weiss, 1974 for details of item development and norming). Each item had been normed on groups of college undergraduates. Norming resulted in estimates of item difficulty (proportion correct), and item discrimination indicated by the biserial correlation of each item with total score on the norming tests. Approximations to the normal ogive item parameters "a" and "b" were determined by the following formulas (Lord & Novick, 1968, pp. 376-378).

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

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

where a is the normal ogive index for discrimination;

b is the normal ogive index for difficulty;

r_b is the biserial correlation coefficient between item response and total score;

f(p) is the inverse of the cumulative normal distribution corresponding to the proportion correct

The item pool was not composed of an equal number of items at each level of difficulty; rather, there were many highly discriminating items which were relatively easy, and fewer highly discriminating items which were difficult.

Construction of the pyramidal tests. Three different pyramidal tests were used in this study. All were 15-stage

fixed branching models with a constant step size. All used an up-one/down-one branching rule (Weiss & Betz, 1973; Weiss, 1974).

For pyramids 1 and 2, the following rationale was used in test construction. Each test was to be administered with a conventional test; therefore, those items used in the conventional test were excluded from the pyramidal tests, in order to avoid a deceptively high correlation between scores from the two testing strategies. This resulted in an important constraint in the construction of the pyramidal tests. Since the conventional test was peaked at b = 0, many highly discriminating items of moderate difficulty were unavailable for the pyramid. However, the pyramidal structure, as illustrated in Figure 1, shows that most items required by this strategy fall into the range of moderate difficulty with fewer items required at extreme levels of difficulty. In general, $n(n+1)/2$ items are required for an n-stage pyramid. Thus, $15(15+1)/2$ or 120 items were needed to build a complete 15-stage pyramidal structure. In order to construct a symmetric pyramid of 15 stages having an initial item of median difficulty and terminal items which ranged in value from -3.0 to +3.0 standard deviations, a step size of b = 0.2 was necessary. That is, increases or decreases in item difficulty from one stage to the next were fixed at a normal ogive difficulty value of 0.2.

Appendix A shows the item difficulty and discrimination structure of the three pyramids used in this study. Tables A-1 and A-2 indicate that the initial item presented to all testees in pyramids 1 and 2 had a difficulty of b =-.05. A correct response branched the subject to a more difficult item at stage 2 (b = .21), while an incorrect response branched him to an item easier than the first (b = -.13). This process was repeated until each subject had attempted 15 items. Once the difficulty of the initial item and the step size had been determined, the remaining items in the pool were divided into 29 groups, with all items in a group having about the same "b" value and an "a" value of .30 or higher. These groups correspond to the 29 columns of items in the tables of Appendix A.

It has been suggested by Paterson (1962) that within each column items be ordered according to discrimination, with the most discriminating item appearing first. In pyramids 1 and 2, there are several exceptions to this rule, as shown in Tables A-1 and A-2. For example, in column 18 of Tables A-1 and A-2, the second item is the one with the highest discrimination. Similarly in column 16 of these tables the best discriminating item is fourth, not first. In constructing these two pyramids, in cases in which the difficulties of items varied widely within a column, item

difficulty was considered more important than item discrimination. Pyramid 3 was structured so that item discriminations were ordered from highest to lowest within each column (see Table A-3).

For the first group of subjects, the pyramid 1 test was presented with a 40-item conventional test. After the initial administration, two errors were found in the pyramidal test; items of inappropriate difficulty were located in difficulty level 12 at stages 4 and 6. Because both items appeared in early stages of the structure, many testees (about one-third of the group) attempted one or both of them. Pyramid 2 was a modified version of the first pyramid, with the errors corrected. Half the subjects received the original pyramid on retesting and the remaining subjects completed the modified version in order to see whether errors in test construction would significantly affect results.

Pyramid 3 (Appendix Table A-3) was administered to a separate group of testees several months after the first two pyramids had been administered. This pyramid was to be given with other adaptive tests which used large numbers of items from the vocabulary pool. Thus, no attempt was made to exclude any items from the pyramid. Since a greater number of highly discriminating items of median difficulty were available, and since items were ordered within a column solely on the basis of their discriminations, the average item discrimination for this test was higher than that of pyramids 1 and 2.

Table 1 presents means and standard deviations for the difficulties, discriminations, and step sizes of the three pyramidal tests. As Table 1 shows the three pyramids are essentially equivalent with respect to mean difficulities of the items (although pyramid 3 is slightly easier than the other two), mean item discriminations (although pyramid 3 has items of slightly higher discriminations), variability of both item difficulties and discriminations, and average step size. Pyramid 1 has considerably larger variability of step size than do pyramids 2 or 3, due solely to the effect of the two items of inappropriate difficulty present in pyramid 1.

Construction of the conventional test. The conventional test used in the study was a peaked test composed of 40 items. Items with p-values of about .60 and high biserial correlations were selected from the item pool. Appendix Table A-4 presents the normal ogive difficulty and discrimination parameters for each item in the conventional test. Table 1 shows means and standard deviations of these normal ogive parameters for both the difficulty and discrimination of the conventional test. As Table 1 indicates, the mean difficulty

Table 1

Summary of Normal Ogive Parameters for Pyramidal and Conventional Tests

Test	Difficulty (b)		Discrimination (a)		Step Size	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Pyramid 1	-.023	1.326	.725	.450	.212	.263
Pyramid 2	-.002	1.306	.728	.450	.208	.079
Pyramid 3	-.094	1.256	.799	.457	.199	.080
Conventional	-.188	0.586	.543	.111	--	--

of the conventional test (-.188) was lower than that of any of the pyramids. The conventional test was constructed to adjust its average difficulty for guessing (Betz & Weiss, 1973, p. 15). On the other hand, the mean difficulty of the pyramids was set at the mean difficulty of the group being measured. The pyramid was not adjusted for guessing since it was assumed that, as a result of the adaptive test's capacity to adjust difficulty level to the individual's ability, guessing was less likely to occur (Weiss & Betz, 1973). Since the conventional test was a "peaked" test, the standard deviation of its difficulties was considerably less than that of the pyramidal tests, which were constructed to measure along an ability continuum.

Table 1 also shows that the adaptive tests were composed of more discriminating items than the conventional test. The latter test was constructed to approximate the conventional tests used in Lord's (1970, 1971a,b) studies (see Weiss & Betz, 1973). It has been suggested, however, that adaptive tests require more highly discriminating items to be effective (e.g., Urry, 1970). Thus, the pyramidal tests used the most discriminating items available in the item pool, within the limitations of the difficulty structure required. This latter fact accounts for the larger variability of discrimination indices for the pyramidal test as compared to the conventional test.

Scoring the Pyramidal Tests

Six scoring methods were used to estimate ability in order to determine which provided the most accurate and most stable estimates. Method 1 is the simple number correct score which has been used by Lord (1970, 1971a,b). For a 15-stage pyramid, sixteen different number correct scores are possible (0 to 15). Method 2 involved computing the mean difficulty of all items attempted for each subject. Lord (1970, 1971a,b) has suggested a similar approach in which the first item is omitted and an hypothetical 16^{th} item is included. Method 3 is analogous to the second; in this method, the mean of the difficulties of the correctly answered items was obtained. In method 4, a subject's score was the difficulty of the final item attempted in the pyramid. Since one objective of adaptive testing is to administer items appropriate to the ability level of the testee, the point at which he/she finishes the test can be considered a good estimate of ability (Lord, 1970). While Bayroff (1960) used the p-value of the final item reached as the testee's score, the normal ogive parameters used in the present investigation are more easily interpretable as an estimate of the subject's ability level.

Method 5 employs an hypothetical 16^{th} item. Since method 4 does not take into account the correctness or incorrectness

of the testee's final response, this method branches the testee to an hypothetical item whose difficulty would be that of the 16^{th} item, were one to be given. Lord (1970, 1971a,b) has called this the "final difficulty score." Values for the $n+1^{th}$ items were computed by averaging the difficulties of all items in its column. Values for the two extreme $n+1^{th}$ items were obtained by using the mean difference between the remaining fourteen items in the $n+1^{th}$ stage and adding it (or subtracting it, in the case of the lower extreme) to the difficulty of the $n+1^{th}$ item adjacent to it.

Scoring method 6 was the all-item score developed by Hansen (1969). In this method, two points are given for a correct answer. In addition, 2 points are added for each item in that stage which is easier than the one attempted, and one point more is added for the next most difficult item in that stage; all more difficult items are scored zero. For an incorrect response, 0 points are given for the item attempted and for all items of greater difficulty in the same stage. One point is added for the next easier item in the same stage, and 2 points are given for all other items of lesser difficulty in the same stage. In this way, all-item scores assign a value to all 120 items in the pyramid for each subject, even though only 15 items were attempted. In contrast to all other scoring methods in which only items actually answered by the testee receive a score, this procedure may provide a method for assessing the internal consistency reliability of pyramidal tests by standard reliability formulas. Scores for this method ranged from 0 to 240.

Test Administration and Subjects

Both conventional and pyramidal tests were administered by cathode-ray-terminals (CRTs) acoustically coupled to a time-shared computer. Items were presented on the CRT screen and testees indicated their response by typing in the number of the correct alternative to the multiple-choice item. Following their response, the next item appeared on the screen. Since the first item of the second test appeared immediately after the final item of the first test, subjects were not aware that two tests were being given (see DeWitt & Weiss, 1974, for details of the computer system controlling test administration).

Subjects were all undergraduates enrolled in either general psychology or psychological measurement and statistics courses at the University of Minnesota. None had any previous experience with computerized testing. Instructional screens explaining the operation of the CRTs were provided prior to testing and a proctor was present in the testing room to provide further assistance to any

testee having difficulty with the equipment. Testees were permitted as much time as necessary to complete the tests and were so informed before the tests were begun.

For the Pyramid 1 study, 250 subjects were originally tested with both the pyramidal and conventional tests. One hundred twenty-five subjects completed the pyramidal test first and the remaining 125 were given the conventional test first. Each subject was retested about seven weeks later. The mean interval between test and retest was 52.5 days; the standard deviation was 7.5 days, and retest intervals ranged from 39 to 70 days. At retest, the group was randomly divided into two subgroups; half the subjects received a retest of pyramid 1 plus a numeric norming test (N=101); while the remaining half was administered the revised pyramid, pyramid 2, and the same conventional test (N=103). Thus, subgroup 1 yielded test-retest data on pyramid 1, while subgroup 2 yielded retest data on the conventional test and an approximation to an alternate form retest for pyramids 1 and 2.

Pyramid 3 was administered with a stradaptive test (Weiss, 1973) to 142 testees. On retest, 138 subjects were administered the same pyramid and a two-stage test. In both administrations, the order of test presentation was randomized. Complete test-retest data on pyramid 3 was available for 128 subjects. The test-retest interval for pyramid 3 was also about 7 weeks with a mean of 49.2 days, a standard deviation of 4.8 days, and a range of 40 to 63 days.

Analysis

The general outline for the studies using each of the pyramidal tests is shown in Table 2. The data to be analyzed in the Pyramid 1 study consisted of two sets of six pyramidal scores, one set for the initial test and one for the retest. Scores for the conventional test (number correct) were available only for the initial test on this group. Those testees completing Pyramid 1 at time 1 and Pyramid 2 at time 2 also had two sets of six scores. Conventional test scores were available for both test administrations. Thus, for this group the test-retest stabilities of the pyramidal test could be compared with that of the conventional test. No conventional test was administered with Pyramid 3. Subjects completing this test at initial testing and at retest were scored by the same six methods used for the other pyramidal tests.

Thus, the design permitted analysis of the stability of scores on pyramid 1 (group 1), stability of scores on a

Table 2

Design for Analyses of Pyramids 1, 2, and 3

Group	Time 1 Tests Administered	N	Time 2 Tests Administered	N
1	Pyramid 1 and Conventional Test	125	Pyramid 1 and Numeric Test	112
2	Pyramid 1 and Conventional Test	125	Pyramid 2 and Conventional Test	112
3	Pyramid 3 and Stradaptive Test	142	Pyramid 3 and Two-Stage Test	138

pyramidal test with revisions in the item structure (pyramids 1 and 2 on group 2), test-retest stability for the conventional test (group 2), and stability of a pyramidal test (pyramid 3) constructed differently than the other pyramids (group 3).

Order effects. To determine whether order of administration significantly affected test scores, the 250 testees who completed pyramid 1 at time 1 (groups 1 and 2) were randomly divided into two subgroups. The pyramidal test was administered first and the conventional second to 125 testees. The order was reversed for the remaining 125. In this way, fatigue or practice effects or carry-over effects between strategies could be detected. T-tests were used to determine whether the differences between the mean scores for each order were statistically significant for the initial test administration. Subjects administered Pyramid 3 were divided into two subgroups on both test and retest. The first was given the pyramidal test first and a stradaptive test (Weiss, 1973) second. The order was reversed for the remaining subjects. Since a different adaptive test (a two-stage test; Betz & Weiss, 1973) was administered with pyramid 3 during the retest, testees were again divided into two groups with respect to order of administration, and t-tests computed for each scoring method.

Score distributions. Two previous empirical investigations using pyramidal testing models have found that score distributions have been negatively skewed, with many testees obtaining near maximum scores. Seeley, Morton, and Anderson (1962) reported that such a result could be attributed either to the scoring method used or the difficulty of the test. Bayroff and Seeley (1967), using two 8-stage pyramidal tests, found scores distributed approximately normally for the verbal section but negatively skewed for both the numerical section and the conventional test. Hansen (1969) however, found that for one scoring method, a more rectangular distribution of scores was obtained with pyramidal tests than with conventional tests.

One objective, then, of the present study was to investigate the distributions of scores on the 40-item conventional test and those derived from each pyramidal scoring method. These analyses were designed to examine (1) the appropriateness of test difficulty, (2) the relative variabilities of each of the various scoring procedures, and (3) the shape of the obtained score distributions.

In order to express the variability of the pyramidal scoring methods in a common unit, the standard deviations for each scoring method were divided by the range of potential scores and the results expressed as the "proportion

of range utilized" (Betz & Weiss, 1973). The ranges for each scoring method were determined in the following manner: (1) the "number correct" range was simply 15-0 = 15 for all three pyramids; (2) the "mean difficulty of all items attempted" range was obtained by subtracting the mean difficulty score made by a testee answering all items incorrectly from the score of one responding correctly to all 15 items; (3) the "mean difficulty of all items correct" range was obtained by subtracting the lowest possible $N+1^{th}$ score from the "mean difficulty" score of a testee with 15 correct responses; (4) the "final item difficulty" range was the difference between the easiest and most difficult terminal items while (5) the "$n+1^{th}$ item difficulty" range was the difference between the two extreme $n+1^{th}$ values; and (6) the all-item score range was 240 for all three pyramids. Exact values for these ranges are summarized in Appendix B.

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

Stability. Previous investigations of pyramidal testing have usually been concerned with the correlation between a short branching test and a longer conventional test. None have studied the relative stabilities or internal consistency reliabilities of conventional versus pyramidal tests. To investigate the accuracy of each scoring method, test-retest correlations were computed for all testees completing both administrations of the pyramidal and conventional tests. In order to detect curvilinear relationships in test-retest stability, eta coefficients were also computed and each bivariate relationship was tested for curvilinearity (McNemar, 1969, pp. 315-317). These data were expected to yield initial information on the relative utility of the various scoring methods for making longitudinal predictions.

To evaluate the effects of the length of the time interval between test and retest on stability, subjects completing both tests were divided into three groups. The first was composed of those testees whose test-retest interval was short (39 to 49 days for pyramids 1 and 2, 40 to 46 days for pyramid 3); the next group had a moderate test-retest interval (50 to 58 days for pyramids 1 and 2, 47 to 53 days for pyramid 3) and the last had the longest interval (59 to 70 days for pyramids 1 and 2, 54 to 63 days for pyramid 3). Test-retest correlations were then calculated separately for each group. Both the time interval and the number of subjects were kept approximately equal for each pyramid and the conventional test.

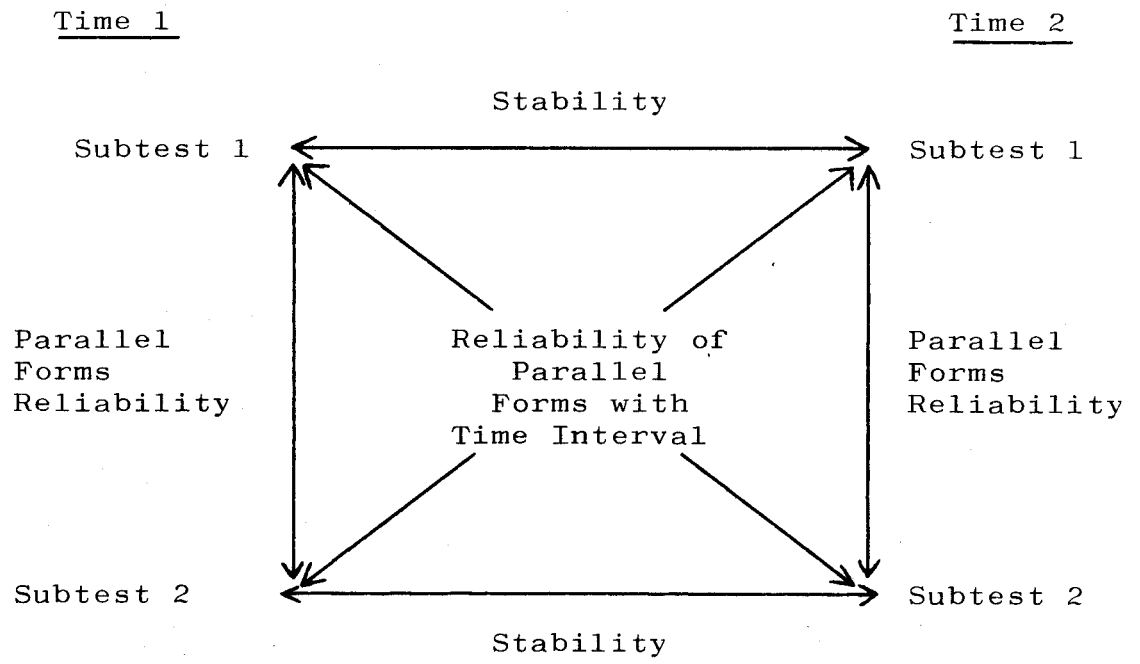
Memory effects. Stability is affected by memory. When a conventional test is administered to the same subjects twice, test-retest correlations may be spuriously high because subjects may remember how they answered items the first time and respond in the same way on second testing. For a pyramidal test, however, subjects may be administered a different set of items during the retest if they move through the pyramidal structure through pathways different from those taken during the original test. Thus, it is possible for subjects completing the same pyramidal test twice to obtain the same score both times, while repeating considerably fewer items than would be the case for a conventional test. For this reason, memory effects are likely to be smaller in pyramidal tests, and test-retest correlations may not be as inflated by memory effects as those for a conventional test of comparable length.

In order to evaluate the effects of memory, the 40-item conventional test was divided into two 15-item parallel subtests. The shortened conventional subtests were comprised of only 15 items to facilitate comparison to the 15-stage pyramidal tests. The following method was used. A bivariate graph was constructed with item difficulty on the abscissa and discrimination on the ordinate. The 40 items were plotted, and the fifteen pairs of items whose "a" and "b" values most nearly matched were selected. Members of each pair were randomly assigned to each of the two parallel subtests. Item parameters for the items of both parallel subtests are given in Appendix C. As Appendix C shows, the two subtests could be considered parallel since the means and standard deviations of both their difficulty and discrimination parameters were almost identical.

Figure 2 indicates diagrammatically the design for the analysis of memory effects. The degree of similarity between the two parallel forms of the 15-item conventional subtests at each of the test administrations is indicated by the two vertical lines; these are parallel forms reliability coefficients. The horizontal lines represent the test-retest stability correlations for both 15-item subtests. Because all 15 items are repeated this condition allows the maximum effect for memory. The diagonal lines show the correlations between different 15-item subtests at different times. If memory effects were present these correlations should follow a specified pattern. First, since subjects attempt the same items twice, the stability correlations should be the highest in the analysis. Secondly, these test-retest correlations would be higher for either subtest than the correlation between one subtest at time 1 and the other at time 2

Figure 2

Design for the analysis of memory effects in the conventional test



since testees would have attempted identical items within forms and completely different items across forms. The latter correlations represent a "no memory" condition. These should be the lowest in the analysis, as memory effects would not be present and a time interval separates the two test administrations. Finally, the parallel forms correlations, which involve no repeated items and, therefore, no memory effects should fall intermediate between the memory condition (stability correlations) and the no memory condition (parallel forms with time interval).

On the pyramidal test, most testees could be expected to attempt an intermediate number of identical items on test and retest. Therefore, it would be expected that stability estimates of the pyramidal test would fall between the extremes of the "no memory plus time interval" and "maximum memory" conditions for the conventional 15-item subtests described above, if the stability of the pyramidal testing strategy did not differ substantially from that of a conventional test of the same length.

Change Analysis. When a conventional test is administered to the same subjects more than once, memory and practice effects may operate to increase retest scores. No investigation has yet attempted to find similar effects in adaptive testing. In order to determine whether scores on the conventional and pyramidal tests changed significantly from one testing to the next, correlated t-ratios were computed contrasting mean scores for the initial and retest administrations. These analyses were conducted for each method of scoring the pyramidal tests and for each pyramid, to determine whether scoring methods and/or the structure of the pyramid had differential effects on mean score changes.

Internal Consistency Reliability. Measures of the internal consistency reliability of both the conventional and pyramidal tests were obtained by the Hoyt (1941) method. In order to compute such an index, a score for every subject on each item must be computed. As testees completed only a small fraction of the total number of items in the pyramidal tests, estimates of the probable scores on unattempted items were made according to the procedures of the "all-item" scoring method described above. The Spearman-Brown formula was used to equate the number of items between the conventional and pyramidal tests since the pyramidal test using the "all-item" score had three times the number of items as the conventional test. Hansen (1969) employed a similar method for obtaining the KR-20 reliability indices for a number of four-stage pyramidal tests.

Table 3

Means and Standard Deviations for Two Subgroups Completing Pyramid 1, Time 1, Pyramid 3, Time 1 and Pyramid 3, Time 2

Scoring Method	Pyramid 1, Time 1					Pyramid 3, Time 1					Pyramid 3, Time 2				
	Pyramid First (N=125)		Pyramid Second N=125)			Pyramid First (N=73)		Pyramid Second (N=69)			Pyramid First (N=71)		Pyramid Second (N=67)		
	Mean	S.D.	Mean	S.D.	t(248df)	Mean	S.D.	Mean	S.D.	t(140df)	Mean	S.D.	Mean	S.D.	t(136df)
Pyramid															
Number Correct	8.14	2.33	7.67	2.62	1.48	8.03	2.67	7.35	2.30	1.63	8.30	2.68	7.72	2.35	1.35
Mean difficulty of all items attempted	.08	.63	-.01	.66	1.05	.02	.57	-.12	.51	1.57	.11	.58	-.08	.53	1.96*
Mean difficulty of all items corrected	-.06	.70	-.17	.76	1.21	-.08	.59	-.23	.54	1.62	-.00	.62	-.19	.56	1.91
Difficulty of final item	.21	.97	.05	1.06	1.27	.04	1.00	-.20	.83	1.54	.21	.99	-.04	.90	1.54
Difficulty of $N+1^{th}$ item	.23	1.02	.02	1.14	1.54	.12	1.07	-.15	.92	1.62	.23	1.06	.00	.94	1.37
All item score	130.77	48.57	122.82	51.77	1.25	129.80	49.99	117.44	44.94	1.55	136.73	50.34	121.97	46.04	1.80
Conventional[a] Number Correct	22.93	8.64	22.53	8.38	0.37										

[a]For the conventional test the order of administration is opposite to that of the pyramidal test.
*Statistically significant at $p \leq .05$

Relationships among scoring methods. To determine which pyramidal scoring methods were most similar and which was most highly related to conventional test scores, each score was correlated with every other score. Both the product moment correlation and the correlation ratio were computed for each pair of scores.

Results

Order Effects

Table 3 provides results of the analyses of the effects of order of administration on scores for pyramidal and conventional tests. Means and standard deviations for both groups completing each test first or second in a paired administration are given for each method of scoring the pyramidal tests and for the conventional test. Of 19 t-tests, only one of the t-ratios for the difference between the mean scores for each order was statistically significant at the .05 level. There was, however, a trend showing that when any one of the three pyramidal tests was administered first, subjects tended to make slightly higher mean scores than those who attempted that test second. For the conventional test, mean score differences were also not statistically significant, but the slight difference in means was in the opposite direction. Since order did not appreciably affect scores on the pyramidal or conventional tests all subsequent analyses combined the data from the two order groups.

Score Distributions

Pyramidal tests. Table 4 shows descriptive statistics for the first administration of pyramid 1 and the conventional test. Similar data is shown for pyramids 2 and 3, and for the retests of all pyramids and the conventional test, in Appendix D. Mean scores shown in Table 4 for both tests indicated that, on the average, the testees answered approximately half (7.90) of the fifteen items in the pyramid correctly, suggesting that the difficulty of the test was appropriate for the ability of the subjects tested. This result was also found for the pyramid 1 retest (Appendix Table D-1), the pyramid 2 administration (Table D-2) and both administrations of pyramid 3 (Tables D-3 and D-4). As might be expected, the mean difficulty score for all items attempted (.04) was higher than the mean difficulty score for all items answered correctly (-.12), indicating that testees usually responded incorrectly to those items which were above their ability level.

Table 4

Descriptive Statistics for the First Administration of Pyramid 1 and Conventional Test (N = 250)

Test and Scoring Method	Mean	Median	S.D.	Proportion of Range Used	Skew	Kurtosis
Pyramidal						
Number Correct	7.90	7.73	2.49	.17	.27	-.45
Mean difficulty of all items attempted	.04	.01	.65	.22	.08	-.96*
Mean difficulty of all items correct	-.12	-.06	.73	.16	-.21	-.82*
Difficulty of final item	.13	.01	1.02	.18	.14	-.61
Difficulty of N+1th item	.12	.09	1.08	.17	.25	-.70*
All-item score	126.80	122.00	50.25	.21	.16	-.90*
Conventional						
Number Correct	22.73	21.70	8.50	.21	.12	-.96*

*Statistically different from zero kurtosis at $p \leq .05$.

Standard deviations for all methods of scoring the first administration of pyramid 1 are given in Table 4. Because the scoring methods used for the pyramidal tests were all on different scales, the variabilities associated with each method are not directly comparable. Thus, Table 4 shows the standard deviations expressed as a proportion of each scoring method's potential range. Inspection of Table 4 (and the supplementary data in Appendix D) indicates that two pairs of scoring methods provided almost identical values for all pyramidal tests. The number correct score and the $n+1^{th}$ scoring method both used from 16 to 19 percent of the possible range. The mean difficulty of all items attempted and the all-item scoring methods used from 19 to 23 percent of the possible range. Expressed as relative variabilities, the mean difficulty of all items attempted and the all-item score had the highest variabilities of the pyramidal scoring methods (.22 and .21 in Table 4). The mean difficulty of all items correct scoring method was lowest in relative variability for pyramid 1. This finding was consistent across all pyramids and all administrations (see Appendix D). Thus, the mean difficulty of all items attempted and the all-item score seem to provide the greatest potential for inter-individual discrimination.

For five of the scoring methods used in the pyramid 1 study, score distributions tended to be positively skewed but not significantly so. Only the mean difficulty of all items correctly answered had a slightly negatively skewed distribution (see Table 4). Both trends were also observed for the retest of pyramid 1 (Table D-1) and for pyramid 2 (Table D-2). All score distributions for pyramid 3 were positively skewed (Tables D-3 and D-4) both on initial test and retest. However, for pyramidal 3, using several of the scoring methods, the degree of skewness indicated a statistically significant departure from normality.

Distributions of scores for four scoring methods for pyramid 1 were highly platykurtic, as shown in Table 4. However, only two scoring method distributions remained significantly platykurtic on retesting (Table D-1). The all-item method of scoring, and the mean difficulty of all items attempted method consistently yielded the flattest distributions. This finding is in accord with the finding of greater relative variability for these methods of scoring the pyramidal test. Results obtained for the pyramid 2 administration (Table D-2) were similar to those for the pyramid 1 retest, with all scoring methods producing

platykurtic distributions and the same two scoring methods showing significant departures from normal kurtosis.

For pyramid 3 the tendency for flat distributions was still present but to a lesser degree (Tables D-3 and D-4). Only the all-item scoring method for the initial administration was significantly platykurtic.

Conventional test. As Table 4 shows, the mean score for the first administration of the conventional test was 22.73. Since the test was composed of 40 items and guessing was possible, this mean score was appropriate, indicating that the test was peaked at the difficulty level of the group being tested. Retest scores (Table D-2) had a mean of 23.40.

The variability of scores on the conventional test, expressed as the proportion of range index, was similar to that of the better pyramidal scoring methods. On retest (Table D-2) the two best pyramidal scores utilized more of their potential range (.23) than did the conventional test (.21). Further, there was a slight, but non-significant, tendency for scores on both administrations of the conventional test to be positively skewed. The score distribution for the conventional test was highly platykuric for the first administration, indicating a statistically significant difference from normality. The distribution remained platykurtic on retesting but was not significantly different from a normal distribution.

Test-Retest Stability

Pyramidal tests. The stability data for the pyramidal tests in Table 5 permit a comparison of the relative stabilities of the various methods for scoring pyramidal tests. For the pyramid 1/pyramid 2 data, three scoring methods yielded substantially lower stabilities. These methods were number correct, difficulty of the $n+1^{th}$ item, and difficulty of final item. This pattern of results was also observed for the pyramid 3 retest and the pyramid 1 retest, using the eta coefficients. It is interesting to note that two of these least reliable scoring methods were among those used by Lord (1970, 1971b) in his theretical studies of pyramidal tests. The most stable scoring methods for scoring the pyramids were the all-item score and the mean difficulty of all items attempted score. Based on the test-retest eta coefficients, mean difficulty of all items correct was consistently the third most stable scoring method but was substantially lower than the other two in the pyramid 1 retest analysis.

Table 5

Test-retest Stabilities of Pyramidal and Conventional Tests

Test and Scoring Method	Pyramid 1 (Time 1) and Pyramid 2 (Time 2) (N=103)		Pyramid 1 Test-Retest (N=101)		Pyramid 3 Test-Retest (N=128)	
	r	eta	r	eta	r	eta
Pyramidal Test (15 stages)						
Number correct	.85	.87	.80	.81	.84	.85
Mean difficulty of all items attempted	.89	.92	.84	.89	.86	.90*
Mean difficulty of all items correct	.87	.91*	.79	.85	.82	.90**
Difficulty of final item	.83	.84	.81	.83	.83	.85
Difficulty of N+1th item	.85	.87	.80	.81	.84	.85
All-item score	.89	.92*	.86	.89	.86	.90*
Conventional Test						
15 Repeated items (maximum memory effect)						
Subtest 1	.88	.89				
Subtest 2	.85	.89*				
15 Different items (no memory effects)						
Subtest 1, Subtest 2	.71	.75				
Subtest 2, Subtest 1	.78	.80				
40 items	.92	.93				

*Curvilinearity statistically significant at $p \leq .05$.
**Curvilinearity statistically significant at $p \leq .01$.

In general, pyramid 1 was the least stable of the pyramids, yielding results substantially below the retest using the corrected pyramid 2. This was probably due to the errors in the construction of pyramid 1 which introduced error into the scores on both testings. Pyramid 3 was slightly less stable than the pyramid 1/pyramid 2 administration. The differences might be attributable to the differences in construction of pyramid 3, or to characteristics of the subjects.

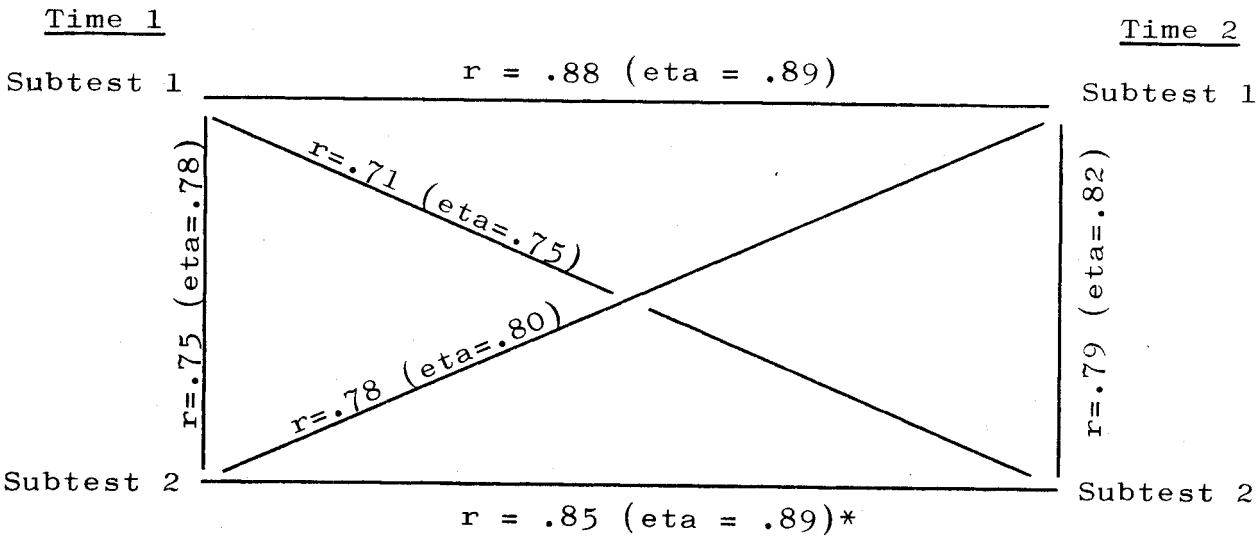
Conventional test. Table 5 also shows the test-retest reliability coefficients for the 40-item conventional test based on the same 103 subjects who completed pyramids 1 and 2. The stability for the conventional test was r=.92 (eta=.93). These were higher than any of the corresponding stabilities for the 15-stage pyramidal tests. However, a comparison of the eta coefficients for the two testing strategies shows that the pyramidal test, composed of only 37.5% of the number of items in the conventional test, was able to achieve stability coefficients not significantly different from those of the conventional test. Both the all-item score and the mean difficulty of all items attempted score yielded test-retest eta coefficients of .92, and the mean difficulty of all items attempted score achieved an eta stability of .91. These compared favorably to the 40-item conventional test stability of .93 for the same subjects. It should also be pointed out that the pyramidal data were based on a modified pyramid at retest (pyramid 2) making the stability correlations not pure test-retest correlations for the pyramidal tests.

Stability comparison. A valid comparison of the relative stabilities of the conventional and pyramidal tests was based on the analysis of memory effects for conventional tests of length equal to that of the pyramidal tests. The memory analysis was based on the assumption that subjects completing the 15-stage pyramidal tests on both test and retest would not attempt the same 15 items on each administration. For the 101 examinees completing pyramid 1 both times, the mean number of items in common was 8.17 with a standard deviation of 3.67. Only five subjects followed the same pathways through the pyramid on both administrations (i.e., answered the same fifteen items both times). The mean number of items in common for the 103 testees in the pyramid 1/pyramid 2 group was 8.25; the standard deviation was 3.47, and three subjects used the same pathways on both administrations.

The test-retest correlations for both 15-item parallel conventional subtests and the correlation of one form with the other across time are presented in Table 5 and summarized in Figure 3. These data serve as a basis for comparison of

Figure 3

Test-retest stability, parallel forms reliabilities and parallel forms stabilities for two 15-item parallel conventional tests (N=103)



*Curvilinearity statistically significant at p = .015.

the 15-stage pyramid and 15-item conventional tests. It can be seen that when the same items were presented at both test and retest, scores were more highly correlated (r=.88 and .85) than when scores on one form were compared with scores on the other during retest (r=.78 and .71) or when scores on different parallel forms were correlated at the same administration (r=.75 and .79). These results are in accordance with the results predicted above and therefore are consistent with an hypothesis that memory effects were operating in the conventional test to inflate test-retest reliability coefficients. Thus, if the pyramidal tests (with an average of only about half the number of items repeated in comparison to the conventional tests of equal length) had stabilities equal to those of the conventional tests, their stability coefficients should lie between the "no-memory" results and the "maximum memory" results.

The data in Table 5 show that three methods of scoring the pyramidal test (the two mean difficulty scores and the all-item score) yielded stability coefficients which were comparable to those of conventional subtest 1 and greater than those of conventional subtest 2 (i.e., maximum memory effects). All pyramidal scoring methods showed higher stabilities than the "across forms" correlations of the parallel conventional tests (no memory effects). Thus, when the effects of memory are taken into account, the pyramidal testing strategy shows greater stability than a conventional test of the same length.

A comparison of the eta coefficients in Table 5 supports the conclusion that the pyramidal test yields more stable scores than the conventional test. Three methods of scoring the pyramidal tests yielded eta stabilities (.92, .91, .92) higher than those of either of the two conventional subtests (.89). This finding is especially significant in that the conventional subtests allowed the possibility of maximum memory effects while the pyramidal test permitted an average of only half the potential for memory effects to operate.

Since the pyramid 1 and pyramid 3 retests used different subjects than the retests of the conventional tests, a direct comparison is not completely appropriate. However, it is interesting to note that even under these circumstances, the best methods of scoring the pyramidal tests yielded eta stabilities equal to or greater than those for the conventional tests with maximum memory effects.

The finding that the stability analysis of the two conventional subtests followed a pattern consistent with the hypothesis of memory effects inflating test-retest reliability coefficients also suggests that the stability coefficients for the 40-item test are inflated by memory effects. From this perspective, the retest eta coefficient of .93 for the 40-item conventional test (with 40 items repeated at retest for all testees) compares very unfavorably with the retest eta of .92 for the retest of the pyramidal test (with an average of 8.25 items repeated).

Retest interval. Table 6 presents the test-retest correlations for the conventional and pyramidal tests as a function of the time interval between administrations. In general, there was little systematic variation in stability with respect to time interval for either the pyramidal or conventional tests. When subjects completed pyramid 1 at time 1 and pyramid 2 at time 2, the medium time interval showed the greatest stability. For Pyramid 1, the short and medium time intervals showed similar stabilities while the long time interval had higher correlations under each method of scoring. For pyramid 3 the highest test-retest correlations were obtained for the short and long time intervals. No general trend is apparent for the 40-item conventional test.

As shown in Table 6 the test-retest correlations for both 15-item conventional subtests were higher than those for the pyramidal tests for the short and long time intervals. For the medium time interval the two mean difficulty scoring methods and the all-item scoring method showed higher stabilities than either of the shortened conventional tests. All pyramidal scoring methods were more stable than conventional subtest 2 for the medium time interval.

Change analysis. Correlated t-ratios comparing mean scores obtained on both administrations of the tests are presented in Table 7. None of the pyramid 1 change scores were significant at $p = .05$, and only the mean difficulty scoring methods showed significant increases when mean scores for pyramid 1 (time 1) and pyramid 2 (time 2) were compared. The latter result is most likely due to the modifications made in pyramid 2 to correct the two items of inappropriate difficulty found in pyramid 1, since the mean difficulty scores would be most affected by this change.

Table 6

Test-retest Correlations as a Function of Time Interval for Pyramidal and Conventional Tests

Test and Scoring Method	Pyramid 1 vs. Pyramid 2 (N=103) 40-49 days N=33	50-58 days N=43	59-70 days N=27	Pyramid 1 (N=101) 39-49 days N=36	50-58 days N=44	59-66 days N=21	Pyramid 3 (N=128) 40-46 days N=42	47-53 days N=62	54-63 days N=24
Pyramidal Test									
Number correct	.84	.88	.82	.79	.78	.87	.86	.81	.90
Mean difficulty of all items attempted	.84	.92	.87	.81	.82	.93	.89	.81	.90
Mean difficulty of all items correct	.82	.90	.85	.77	.74	.91	.89	.75	.88
Difficulty of final item	.79	.86	.83	.78	.81	.91	.84	.82	.88
Difficulty of N+1th item	.85	.87	.82	.82	.78	.84	.85	.82	.90
All-item score	.85	.92	.88	.83	.84	.94	.90	.82	.90
Conventional Test									
15-item form 1	.86	.89	.88						
15-item form 2	.86	.83	.89						
40-items	.94	.91	.93						

Table 7

Correlated t-tests for Pyramidal and Conventional Tests, Time 1 vs. Time 2

Test and Scoring Method	Pyramid 1/Pyramid 2 (N=103)				Pyramid 1 (N=101)				Pyramid 3 (N=128)			
	$\bar{X}_2-\bar{X}_1$	S.E.	t	p	$\bar{X}_2-\bar{X}_1$	S.E.	t	p	$\bar{X}_2-\bar{X}_1$	S.E.	t	p
Pyramidal test (15 stages)												
Number correct	0.05	.147	.33	.742	0.03	.152	.20	.845	0.34	.126	2.67	.008
Mean difficulty of all items attempted	0.09	.031	2.92	.004	0.04	.035	1.21	.229	0.05	.026	1.96	.052
Mean difficulty of all items correct	0.12	.037	3.16	.002	0.05	.044	1.04	.303	0.04	.030	1.44	.152
Difficulty of final item	0.03	.064	.47	.642	0.01	.058	.24	.813	0.16	.047	3.31	.001
Difficulty of N+1th item	0.06	.061	1.05	.296	0.00	.067	.01	.996	0.14	.050	2.75	.007
All-Item Score	3.11	2.509	1.24	.218	2.10	2.561	.82	.414	5.10	2.182	2.34	.021
Conventional Test (40 items)	1.04	.34	3.05	.003								

For pyramid 3, the probabilities associated with four of the six scoring methods were less than .05. Thus, the increases from time 1 to time 2 were statistically significant, in one case at the .001 level. On the conventional test, scores were higher on the retest and the difference was significant at the .01 level.

The significant increases in test scores between test and retest seen in the results for pyramid 3 contrast sharply with the nonsignificant increases for the retest of pyramid 1 and with those for pyramid 1/pyramid 2 retest. The time interval between test and retest was approximately the same for all administrations, so it can probably be ruled out as a cause of this discrepancy. It is possible that characteristics of the subject groups contributed to the difference.

Also, differences in the construction of the pyramidal tests and/or differences in administration could have caused the significant mean differences for pyramid 3. As was indicated earlier, pyramid 3 was constructed using all available items in the item pool regardless of whether they were to be administered under another adaptive strategy, whereas in constructing pyramids 1 and 2 item overlap was avoided. As Table 1 shows, pyramid 3 was first administered with a stradaptive test which had a considerable degree of item overlap with the pyramidal test. As a result, testees would likely be administered a substantial number of common items on first administration. This might result in a greater memory effect on retest than when the testees answered each item only once on first administration, as they did in pyramids 1 and 2.

The very significant increase in mean scores upon retesting for the conventional test is likely to be a function of memory and/or practice effects, in comparison to the general absence of such effects for the corrected pyramidal retest (pyramid 1/pyramid 2) for the same group of subjects. These results support the memory analyses reported above suggesting that scores on pyramidal tests are less affected by memory than those of conventional tests.

Internal Consistency Reliability

The Hoyt (1941) index of internal consistency reliability for the 40-item conventional test was .89 for the initial administration and .90 for the retest. When the Spearman-Brown correction for triple length was used (in order to make the conventional test comparable to a pyramidal test of 120 items) the reliability increased to .96 for both test and retest. For every administration of the

pyramidal test using the all-item score this index was .99. It would appear, then, that the all-item method of scoring (Hansen, 1969) yields a reliability coefficient which is spuriously high. Such a result may be due to the strong assumptions made about the monotonic relationships of item difficulty and testee response in computing scores under this method. Under this scoring method, error does not affect the items a person does not attempt.

Relationships among Scoring Methods

Table 8 presents the intercorrelation matrix for all pyramidal scoring methods for pyramid 1 and the correlations between pyramidal scoring methods and the conventional test scores. Similar data for the other pyramids are shown in Appendix E.

Pyramidal vs. conventional scores. For pyramid 1 the all-item score correlated more highly (r=.86) with scores on the conventional test than any other scoring method. The mean difficulty of all items attempted scoring method correlated nearly as highly (r=.85) with scores on the conventional test as the all-item score. The same two scoring methods were most highly correlated with the conventional test when pyramid 2 was used (Appendix Table E-1). For both pyramids 1 and 2 the number correct method as well as the $n+1^{th}$ scoring method correlated lower with the conventional test than did the other methods.

Methods of pyramidal scoring. For all test administrations (Table 8 and Appendix E) the highest values obtained in the intercorrelation matrices were those between the number correct and difficulty of the $n+1^{th}$ item scoring methods. Such a correlation should always equal 1.0 as the 16 possible scores for the number correct method (0 through 15) correspond exactly to the scores of the 16 $n+1^{th}$ difficulties, no matter how such difficulties are computed. Lord (1970) has also shown this to be the case. All testees answering a given number of items correctly will be branched to the same $n+1^{th}$ terminal position in the pyramidal structure, regardless of which items were correct. The assumptions needed are that the values for the $n+1^{th}$ scores increase monotonically and that these items are equally spaced on the difficulty continuum. In a properly constructed pyramid this must be the case. However, due to the two item placement errors in pyramid 1 the Pearson correlation between these two scoring methods for pyramid 1 (Table 8) was only .99. For the other pyramidal test (Appendix E) this correlation was 1.0, as would be expected.

Table 8

Intercorrelations of Scores from Pyramid 1 and Conventional Test, Time 1 (N=250)

Test and Scoring Method	Pyramidal Test: Number Correct	Mean difficulty of all items attempted	Mean difficulty of all items correct	Difficulty of Final Item	Difficulty of N+1th item	All-Item Score
Pyramidal Test						
Mean difficulty of all items attempted	r=.92 eta=.92**					
Mean difficulty of all items correct	r=.91 eta=.91*	.99 .99***				
Difficulty of Final Item	r=.98 eta=.98	.94 .95	.93 .95			
Difficulty of N+1th item	r=.99 eta=1.00***	.91 .94	.90 .94***	.97 .97***		
All-Item Score	r=.95 eta=.96**	.99 .99	.98 .99***	.96 .97	.95 .96***	
Conventional Test (40 items)	r=.82 eta=.84*	.85 .90**	.84 .88*	.84 .85	.83 .84	.86 .87

*Curvilinearity statistically significant at $p \leq .05$
**Curvilinearity statistically significant at $p \leq .01$
***Curvilinearity statistically significant at $p \leq .001$

Another strikingly high correlation observed for all administrations of the pyramids was that between the mean difficulty of all items attempted score and the all-item scores (r=.99 for pyramid 1). This high correlation accounts for the fact that stability estimates for these two scoring methods were nearly always equal; stabilities of these scoring methods were always higher than those for any other scoring method. Such a strong relationship might not be expected as the all-item score appears to have only a very approximate relationship with the actual item difficulties. The lowest correlations among pyramidal scoring methods involved the mean difficulty of all items correctly answered score. This finding, in conjunction with the comparatively low stabilities of this scoring method, suggest that it is the least valuable pyramidal scoring method. The mean difficulty of all items correctly answered correlated more highly with the mean difficulty of all items attempted than with any of the other scoring methods. This was expected since both methods involve only simple averaging of the difficulties of some or all of the 15 items administered to an individual. Thus, the mean difficulty of all items correct also correlated highly with the all-item score .

The difficulty of the final item scoring method correlated highest with the $n+1^{th}$ method and total number correct methods. Since, for a certain final item, only two $n+1^{th}$ scores are possible given the structure of the pyramid, such scoring methods will be very highly related. However, the correlations will not be 1.0, since some of the testees answer the final item correctly while others do not.

The all-item scoring method correlated highly with more scoring methods than any other. This finding contrasts sharply with those of Hansen (1969). In that investigation the all-item method had the lowest relationship to the other scoring methods used.

Discussion and Conclusions

The order of test administration was not found to significantly affect mean scores for either pyramidal or conventional tests. The trend for pyramidal test scores to be lower when the pyramid was administered after the 40-item conventional test suggests that fatigue may have affected the testees to some small extent. In a study of two-stage tests, Betz and Weiss (1973) found order effects to be non-significant.

The pyramidal tests used in this study were found to be of appropriate difficulty for the ability of the testees. This is shown by the fact that, for all administrations,

the mean number of items correctly answered was slightly more than half of the total number of items administered. Such results were not obtained by Seeley, Morton and Anderson (1962) in their paper and pencil administration of pyramidal tests. In that case a large percentage of testees obtained the maximum score. This might have been due to the easiness of their test or to the exclusion of many test papers submitted by testees of lower ability who had difficulty in following the branching instructions. When Bayroff and Seeley (1967) administered branched tests by computer, scores on a verbal item pyramid were distributed approximately normally. Thus, it appears that when a good estimate of the general ability level of a group of individuals is known in advance, a pyramidal test of appropriate difficulty can be constructed.

In contrast to the highly negative skew in the Seeley et al. study, distributions for pyramids 1 and 2 were approximately normal with a slightly positive skew for most scoring methods. Only the average difficulty of all items correctly answered score produced a negatively skewed distribution, but again the distribution was approximately normal. For pyramid 3 however, the departure from normality was significant and in a positive direction. This result was unexpected as pyramid 3 was slightly easier than the others.

The trend for most of the pyramidal distributions to be platykutic has been noted by Hansen (1969), who obtained a rectangular score distribution. For pyramids 1 and 2 most of the score distributions were significantly flatter than the normal distribution while for pyramid 3 almost all were not.

The conventional test used in the present study also yielded scores which were significantly platykutic. As Betz and Weiss (1973) have pointed out, this may have been a function of deviations in the peakedness of the conventional test, with a more highly peaked test producing more nearly normal score distributions.

While Betz and Weiss (1973) have found that the two-stage testing strategy yielded scores which utilize a higher proportion of the score range than a conventional test, the pyramidal tests in the present study used a percentage of range equal to or slightly greater than that of the conventional test for only two of the six scoring methods used. These were the mean difficulty of all items attempted and the all-item scores, which were later shown to correlate .99.

A comparison of the scoring methods used for the pyramidal tests indicated that the most stable were the mean difficulty of all items attempted and the all-item scores. Test-retest correlations for these scoring methods approached those of the 40-item conventional test. This finding supports Lord's (1970, 1971b) contention that the average difficulty score is the most appropriate way to score a pyramidal test when the up-one/down-one branching rule is used. In each of the pyramids these scoring methods were consistently more stable than either the difficulty of the $n+1^{th}$ item scoring method, or the number correct score, and they also correlated more highly with conventional test scores than any other scoring methods. One possible explanation for the good results obtained with these two methods is that they utilize more information than the other scoring methods and take account of the different pathways through the test structure. As most of the earlier studies of pyramidal testing (Bayroff, Thomas and Anderson, 1960; Seeley, Morton and Anderson, 1962; Waters, 1964; Bayroff and Seeley, 1967; and Waters, 1970) have used a simple rank ordering of scores essentially equivalent to the number correct score, or $n+1^{th}$ item difficulty score, the correlations with parent tests obtained in these studies might have been higher, had either of the better scoring methods been used.

The time interval between test administrations did not affect the stabilities of either the pyramidal or conventional tests in any consistent manner. But the intervals used were restricted to between six and ten weeks. Longer time intervals would be appropriate to show more clearly whether pyramidal testing provides estimates of abilities which are more stable over time than those of conventional testing.

The analysis of memory effects in the present study indicated that pyramidal testing provides estimates of ability comparable to conventional tests of the same length even though in the conventional test testees attempt the same items at both test and retest, resulting in an inflated estimate of stability due to memory of previous responses. When the effects of memory were controlled for, the pyramidal tests showed higher stabilities than conventional tests with the same number of items.

The analysis of the change in mean scores from test to retest indicated that scores on the conventional test increased significantly. For the pyramidal strategy the significance of increases in test scores depended on the scoring method used and the particular pyramid involved.

For pyramid 1, none of the differences obtained by each scoring method were significant. For pyramid 3, all the differences obtained (except for the two mean difficulty scores) were statistically significant. This result may have been an artifact of the methods of construction and administration of pyramid 3, in contrast to that of the other pyramids. Significant increases in mean scores for the pyramid 1/pyramid 2 administration were found using the mean difficulty scores only. These results were likely due to the errors in the construction of the branching network of pyramid 1.

The internal consistency reliabilities for pyramidal tests obtained by Hansen (1969) for several three- and four-stage pyramids scored by the all-item method were quite high. The present study also obtained extremely high internal consistency reliability for this scoring method. The all-item scoring method, however, makes a strong assumption about the correctness of responses to unattempted items based on actual responses. A correct response to an item is taken as evidence that all easier items in that stage will be answered correctly while almost all more difficult items will be answered incorrectly. Internal consistency reliabilities calculated from such hypothetical response patterns would thus seem to be seriously overestimated. At present, then, the internal consistency reliabilities of adaptive tests would seem to be unmeasurable by conventional methods which require a response to each item by every individual. In one recent study of adaptive testing, Betz and Weiss (1973) were able to measure internal consistency reliability for two-stage tests only by considering the routing and measurement tests as separate conventional tests.

Comparison of the scoring methods used indicated three important facts: (1) the mean difficulty of all items attempted correlates very highly with the all-item score; (2) the number correct and difficulty of the $n+1^{th}$ item scores are also perfectly correlated given a properly constructed pyramid, as has been shown by Lord (1970, 1971b); (3) the all-item score correlates highly with more other scoring methods than any other scoring method. Hansen (1969) found that the all-item scoring method had the lowest over all relationship to three other scoring methods used. This discrepancy may be due to the extremely short tests used in Hansen's study or to the fact that two of Hansen's other scoring methods were not used in the present study.

The major deficiency of the present study was the presence of two errors in pyramid 1. These two items of inappropriate difficulty may have served to increase the mean scores of the pyramid as they were relatively easy items located in positions designed for more difficult items. Seeley, Morton and Anderson (1962) have encountered similar difficulties with their sequential item tests. In that study, "despite repeated checking and cross-checking the ... tests administered in the field showed a number of construction oversights which would require correction before further use could be made of the tests" (Seeley et al., 1962, p. 7). The effects of errors in estimating the difficulties and discriminations of items in pyramidal tests were investigated by Paterson (1962). He found that errors in item difficulty were insignificant when they occurred early in testing. This would seem to indicate that the branching process serves to reduce the effects of items of inappropriate difficulty. As the errors in pyramid 1 were in the fourth and sixth stages, the effects of the errors on the score distribution may have been negligible. The results of the present study support Paterson's finding, however, since the test-retest correlation of scores on pyramid 1 and pyramid 2 were still higher than those of equal length conventional tests when memory effects were taken into account. That these results were obtained from the administration of a pyramidal test with two errors in item placement indicates that pyramidal adaptive tests with errors in their construction will give results similar to those of properly constructed pyramidal tests.

The findings of the present study suggest that pyramidal testing can provide estimates of ability which have stabilities comparable to those of longer conventional tests and greater than those of conventional tests of the same length. Further studies will be needed to determine whether pyramidal testing provides more precise ability estimates throughout the entire range of ability than those of conventional tests and whether pyramidal tests correlate more highly with an external criterion of ability than conventional testing methods.

References

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, 1973.

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

Bryson, R. A comparison of four methods of selecting items for computer-assisted testing. Technical Bulletin STB 72-8, Naval Personnel and Training Research Laboratory, San Diego, December 1971.

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.

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.

Krathwohl, D. R. & Huyser, R. J. The sequential item test (SIT). American Psychologist, 1956, 2, 419.

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. A theory of test scores. Psychometric Monograph, 1952, No. 7.

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. Robbins-Munro 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. & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-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.

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

Mussio, J. J. A modification to Lord's model for tailored tests. Unpublished doctoral dissertation, University of Toronto, 1972.

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.

Stocking, M. Short tailored tests. Princeton, N. J.: Educational Testing Service, Research Bulletin RB-69-63, 1969.

Urry, V. W. A monte carlo investigation of logistic test models. Unpublished doctoral dissertation, Purdue University, 1970.

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

Waters, C. W. Comparison of computer-simulated conventional and branching tests. U. S. Army Behavior and Systems Research Laboratory, Technical Research Note 216, 1970.

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. Strategies of adaptive ability measurement. Research Report 74-x, Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, 1974 (in preparation).

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

Wolfe, J. H. Specification for program BRANCH. Unpublished Memorandum, July 1970.

Wood, R. The efficacy of tailored testing. Educational Research, 1969, 11, 219-222.

Appendix A

Item Difficulty and Discrimination Parameters for Items of the Three Pyramidal Tests and the Conventional Test

Table A-1

Item Parameters for Pyramid 1

Stage	Difficulties (b) and Discriminations (a)																											
1															b= -.05 a= *1.31*													
2															-.13 *.98*	.21 *.86*												
3														-.47 *.71*	-.08 *.91*	.34 *.91*												
4													-1.55 *.77*	-.25 *.86*	.30 *.78*	.65 *.77*												
5												-.73 *.92*	-.40 *.68*	.07 *.76*	.46 *.48*	.79 *.70*												
6											-1.08 *1.23*	-2.19 *.56*	-.21 *.86*	.24 *.66*	.73 *.98*	1.07 *.72*												
7										-1.23 *1.35*	-.75 *.82*	-.40 *.64*	-.09 *.41*	.49 *.56*	.79 *.63*	1.33 *.60*												
8									-1.51 *1.40*	-1.06 *.89*	-.69 *.67*	-.23 *.81*	.14 *1.07*	.65 *.48*	.98 *.52*	1.49 *.62*												
9								-1.68 *1.46*	-1.21 *.91*	-.85 *.75*	-.41 *.59*	.09 *.43*	.46 *.49*	.79 *.45*	1.17 *.52*	1.65 *.39*												
10							-1.87 *1.43*	-1.42 *.92*	-1.07 *.76*	-.58 *.66*	-.29 *.75*	.15 *.97*	.71 *.44*	1.11 *.56*	1.54 *.58*	1.89 *.85*												
11						-2.13 *1.10*	-1.67 *1.02*	-1.31 *.87*	-.81 *.67*	-.40 *.41*	.09 *.37*	.42 *.55*	.75 *.37*	1.30 *.52*	1.61 *.34*	2.03 *.64*												
12					-2.26 *.98*	-1.88 *1.14*	-1.43 *.76*	-1.00 *.67*	-.58 *.48*	-.23 *.43*	.16 *.86*	.65 *.43*	1.01 *.42*	1.40 *.55*	1.93 *.57*	2.31 *.40*												
13				-2.45 *3.00*	-2.08 *.42*	-1.66 *.93*	-1.27 *.56*	-.84 *.65*	-.36 *.40*	.08 *.39*	.37 *.51*	.83 *.37*	1.26 *.39*	1.60 *.35*	2.05 *.49*	2.47 *.47*												
14			-2.72 *3.00*	-2.32 *.80*	-1.92 *1.23*	-1.47 *.67*	-.94 *.60*	-.67 *.30*	-.26 *.40*	.17 *.83*	.62 *.41*	1.01 *.40*	1.44 *.49*	1.79 *.51*	2.27 *.35*	2.67 *.42*												
15	-2.86 *1.01*	-2.45 *3.00*	-2.12 *.41*	-1.65 *.82*	-1.30 *.52*	-.85 *.64*	-.38 *.38*	-.05 *.35*	.33 *.53*	.92 *.37*	1.25 *.43*	1.63 *.32*	2.07 *.43*	2.37 *.34*	2.95 *.84*												
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

Table A-2

Item Parameters for Pyramid 2

Stage	Difficulties (b) and Discriminations (a)																											
1													b= -.05 a= *1.31*															
2													-.13 *.98*		.21 *.86*													
3												-.47 *.71*		-.08 *.91*		.34 *.91*												
4											-.68 *.86*		-.25 *.86*		.30 *.78*		.65 *.77*											
5										-.73 *.92*		-.40 *.68*		.07 *.76*		.46 *.48*		.79 *.70*										
6									-1.08 *1.23*		-.59 *.83*		-.21 *.86*		.24 *.66*		.73 *.98*		1.07 *.72*									
7								-1.23 *1.35*		-.75 *.82*		-.40 *.64*		-.09 *.41*		.49 *.56*		.79 *.63*		1.33 *.60*								
8							-1.51 *1.40*		-1.06 *.89*		-.69 *.67*		-.23 *.81*		.14 *1.07*		.65 *.48*		.98 *.52*		1.49 *.62*							
9						-1.68 *1.46*		-1.21 *.91*		-.85 *.75*		-.41 *.59*		.09 *.43*		.46 *.49*		.79 *.45*		1.17 *.52*		1.65 *.39*					
10					-1.87 *1.43*		-1.42 *.92*		-1.07 *.76*		-.58 *.66*		-.29 *.75*		.15 *.97*		.71 *.44*		1.11 *.56*		1.54 *.58*		1.89 *.85*					
11				-2.13 *1.10*		-1.67 *1.02*		-1.31 *.87*		-.81 *.67*		-.40 *.41*		.09 *.37*		.42 *.55*		.75 *.37*		1.30 *.52*		1.61 *.34*		2.03 *.64*				
12			-2.26 *.98*		-1.88 *1.14*		-1.43 *.76*		-1.00 *.67*		-.58 *.48*		-.23 *.43*		.16 *.86*		.65 *.43*		1.01 *.42*		1.40 *.55*		1.93 *.57*		2.31 *.40*		
13				-2.45 *3.00*		-2.08 *.42*		-1.66 *.93*		-1.27 *.56*		-.84 *.65*		-.36 *.40*		.08 *.39*		.37 *.51*		.83 *.37*		1.26 *.39*		1.60 *.35*		2.05 *.49*		2.47 *.47*
14		-2.72 *3.00*		-2.32 *.80*		-1.92 *1.23*		-1.47 *.67*		-.94 *.60*		-.67 *.30*		-.26 *.40*		.17 *.83*		.62 *.41*		1.01 *.40*		1.44 *.49*		1.79 *.51*		2.27 *.35*		2.67 *.42*
15	-2.86 *1.01*		-2.45 *3.00*		-2.12 *.41*		-1.65 *.82*		-1.30 *.52*		-.85 *.64*		-.38 *.38*		-.05 *.35*		.33 *.53*		.92 *.37*		1.25 *.43*		1.63 *.32*		2.07 *.43*		2.37 *.34*		2.95 *.84*
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

Table A-3

Item Parameters for Pyramid 3

Stage	Difficulties (b) Discriminations (a)																												
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

Table A-4

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

Item Reference No.	b	Ag
58	-.957	.482
221	-.740	.647
307	-.836	.562
386	.136	.697
211	-.720	.609
224	-.785	.543
390	-.731	.627
667	-.726	.568
156	-.631	.647
208	-.681	.582
234	-.687	.512
52	-.282	.606
137	-.739	.400
176	-.897	.338
207	-.526	.602
218	-.928	.332
205	-.618	.472
382	-.481	.638
342	.172	.774
265	.173	.772
645	-.320	.501
661	-.296	.579
670	-.282	.620
327	-.248	.571
50	-.234	.505
144	-.184	.627
369	-.215	.562
233	-.172	.468
139	.189	.417
633	-.078	.501
146	.000	.607
295	-.035	.474
113	.247	.609
267	.188	.436
59	.173	.637
147	1.152	.383
174	1.156	.638
242	.979	.310
306	.969	.490
367	.978	.377

Appendix B

Possible Score Ranges for Three Pyramidal Tests

Scoring Method	Pyramid 1	Pyramid 2	Pyramid 3
Number Correct	15	15	15
Mean difficulty of all items attempted	2.97	2.91	2.79
Mean difficulty of all items correct	4.58	4.58	4.42
Difficulty of Final Item	5.81	5.81	5.48
Difficulty of $N+1^{th}$ item	6.21	6.21	5.98
All-item score	240	240	240

Appendix C

Difficulty (b) and Discrimination (a) Item Parameters for two 15-item Parallel Conventional Subtests

Subtest 1			Subtest 2		
Item reference no.	b	a	Item reference no.	b	a
221	-.740	.647	307	-.836	.562
224	-.785	.543	386	.136	.697
390	-.731	.627	211	-.720	.609
667	-.726	.568	156	-.631	.647
176	-.897	.338	208	-.681	.582
382	-.481	.638	52	-.282	.606
342	.172	.774	207	-.526	.602
670	-.282	.620	218	-.928	.332
50	-.234	.505	265	.173	.772
144	-.184	.627	661	-.296	.579
369	-.215	.562	327	-.248	.571
295	-.035	.474	233	-.172	.468
267	.188	.436	139	.189	.417
59	.173	.637	633	-.078	.501
242	.979	.310	367	.978	.377
Mean	-.253	.554		-.262	.555
s.d.	.505	.124		.500	.118

Appendix D

Descriptive Statistics for Score Distributions of Pyramidal and Conventional Tests

Table D-1

Descriptive Statistics for Retest of Pyramid 1 (N = 112)

Scoring Method	Mean	Median	S.D.	Proportion of Range Used	Skew	Kurtosis
Number correct	8.02	7.43	2.40	.16	.43	-.67
Mean Difficulty of all items attempted	.11	.04	.60	.20	.16	-.99*
Mean Difficulty of all items correct	-.03	-.04	.66	.14	-.13	-.75
Difficulty of final item	.18	-.05	.96	.17	.33	-.70
Difficulty of N+1th item	.16	-.06	1.06	.17	.36	-.89
All-item score	131.26	118.50	47.98	.20	.29	-.97*

*Statistically different from zero kurtosis at $p \leq .05$.

Table D-2

Descriptive Statistics for Second Administration of Pyramid 2 and Conventional Retest (N = 112)

Test and Scoring Method	Mean	Median	S.D.	Proportion of Range Used	Skew	Kurtosis
Pyramid						
Number Correct	7.78	7.68	2.75	.18	.19	-.75
Mean difficulty of all items attempted	.08	.01	.67	.23	.19	-.93*
Mean difficulty of all items correct	-.05	-.08	.72	.16	-.14	-.61
Difficulty of final item	.10	.01	1.13	.19	.20	-.71
Difficulty of N+1th item	.12	.06	1.15	.19	.18	-.77
All-Item Score	126.63	121.50	55.52	.23	.16	-1.00*
Conventional						
Number Correct	23.40	22.50	8.57	.21	.01	-.89

*Statistically different from zero kurtosis at $p \leq .05$.

Table D-3

Descriptive Statistics for First Administration of Pyramid 3 (N = 142)

Scoring Method	Mean	Median	S.D.	Proportion of Range Used	Skew	Kurtosis
Number Correct	7.70	7.15	2.51	.17	.55*	-.39
Mean difficulty of all items attempted	-.04	-.16	.54	.19	.47*	-.80
Mean difficulty of all items correct	-.15	-.25	.57	.13	.31	-.73
Difficulty of Final Item	-.08	-.28	.93	.17	.54*	-.41
Difficulty of $N+1^{th}$ Item	-.01	-.21	1.01	.17	.54*	-.44
All-Item Score	123.79	113.50	47.84	.20	.47*	-.83

*Significantly different from zero skew at $p \leq .05$.
**Significantly different from zero kurtosis at $p \leq .05$.

Table D-4

Descriptive Statistics for Retest of Pyramid 3 (N = 138)

Scoring Method	Mean	Median	S.D.	Proportion of Range Used	Skew	Kurtosis
Number Correct	8.01	7.47	2.53	.17	.45*	-.13
Mean Difficulty of all items attempted	.02	-.12	.56	.20	.43*	-.53
Mean Difficulty of all items correct	-.10	-.22	.60	.14	.11	.01
Difficulty of final item	.09	-.06	.96	.18	.34	-.26
Difficulty of N+1th item	.12	-.08	1.01	.17	.44*	-.23
All-Item Score	129.57	115.83	48.69	.20	.40	-.60

*Significantly different from zero skew at $p \leq .05$.

Table E-1

Intercorrelations of scores from Pyramid 1 Retest (N=112)
in lower triangle and Pyramid 2 and Conventional test at time 2 (N=112) in upper triangle

Test and Scoring Method	Pyramidal Test: Number correct	Mean difficulty of all items attempted	Mean difficulty of all items correct	Difficulty of final item	Difficulty of N+1th item	All item score	Conventional number correct
Pyramidal Test Number correct	r=	.94	.94	.98	1.00	.96	.87
	eta=	.94	.94	.98	1.00	.96	.88
Mean difficulty of all items attempted	r=.91		.99	.96	.94	1.00	.89
	eta=.91		.99	.97	.96	1.00	.92
Mean difficulty of all items correct	r=.90	.99		.95	.94	.99	.89
	eta=.90	.99		.97	.95	1.00***	.91
Difficulty of final item	r=.98	.93	.92		.98	.98	.87
	eta=.98	.94	.94*		.98	.98**	.88
Difficulty of N+1th item	r=.99	.90	.89	.96		.96	.87
	eta=.99***	.92	.93	.97**		.96	.88
All-item score	r=.95	.99	.98	.96	.94		.89
	eta=.95	.99	.99***	.96	.94		.91

*Curvilinearity statistically significant at $p \leq .05$
**Curvilinearity statistically significant at $p \leq .01$
***Curvilinearity statistically significant at $p \leq .001$

Table E-2

Intercorrelations of scores from Pyramid 3, time 1 (N=142) in lower triangle, and Pyramid 3, time 2 (N=138) in upper triangle

	Pyramidal Test					
Scoring Method	Number Correct	Mean difficulty of all items attempted	Mean difficulty of all items correct	Difficulty of final item	Difficulty of N+1th item	All-item score
Number correct	r=	.93	.93	.98	1.00	.96
	eta=	.93	.92	.97	1.00	.96
Mean difficulty of all items attempted	r= .93		.99	.95	.93	1.00
	eta= .93		.99	.95	.94	1.00
Mean difficulty of all items correct	r= .91	.99		.94	.93	.99
	eta= .92	.99		.95	.94	.99**
Difficulty of final item	r= .98	.94	.93		.98	.97
	eta= .98	.95	.95		.97	.97
Difficulty of N+1th item	r=1.00	.93	.91	.98		.96
	eta=1.00	.93	.94	.98		.96
All-item score	r= .95	1.00	.99	.96	.96	
	eta= .96	1.00	.99	.97	.96	

**Curvilinearity statistically significant at $p \leq .01$.