DOCUMENT RESUME

ED 118 602 TM 005 086

AUTHOR Vale, C. David; Weiss, David J.

TITLE A Study of Computer-Administered Stradaptive Ability Testing. Research Report 75-4.

INSTITUTION Minnesota Univ., Minneapolis. Dept. of Psychology.

SPONS AGENCY Office of Naval Research, Washington, D.C. Personnel and Training Research Programs Office.
REPORT NO RR-75-4
PUB DATE Oct 75
NOTE 55p.
AVAILABLE FROM Psychometric Methods Program, Dept. of Psychology, Univ. of Minnesota, Minneapolis, Minnesota 55455 (RR-75-4, free while supplies last)

EDRS PRICE MF-$0.83 HC-$3.50 Plus Postage
DESCRIPTORS *Ability; Ability Grouping; Branching; College Students; Comparative Analysis; *Computer Oriented Programs; Group Tests; Individual Differences; Item Analysis; Psychometrics; Response Style (Tests); Scoring Formulas; *Testing; Vocabulary
IDENTIFIERS *Stradaptive Testing

ABSTRACT

A conventional vocabulary test and two forms of a stradaptive vocabulary test were administered by a time-shared computer system to undergraduate college students. The two stradaptive tests differed in that one counted question mark responses (i.e., omitted items) as incorrect and the other ignored items responded to with question marks. Stradaptive test scores were more consistent with the hypothesized nature of the population distribution of verbal ability. When corrected for differing levels of item discrimination and memory effects, the test-retest stabilities of the two testing strategies were about equal. Scores on one form of the stradaptive test were found to be very stable for testees who had highly consistent response records on initial testing. Stability of "subject characteristic curve" data was high, suggesting the usefulness of these data for describing test-testee interactions. Of the 10 stradaptive ability scores studied, which grouped into four clusters, average difficulty scores had the highest stabilities. Analysis of difficulties of items associated with correct, incorrect, and question mark responses suggested that items with question mark responses should not be ignored, but should be treated as incorrect responses in branching decisions. Suggestions for future research on the stradaptive testing model are made. (Author)

A STUDY OF COMPUTER-ADMINISTERED STRADAPTIVE ABILITY TESTING

C. DAVID VALE

AND

DAVID J. WEISS

RESEARCH REPORT 75-4

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

OCTOBER 1975

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

Unclassified

BEST COPY AVAILABLE

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Research Report 75-4	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A Study of Computer-administered Stradaptive Ability Testing		5. TYPE OF REPORT & PERIOD COVERED Technical Report 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) C. David Vale and David J. Weiss		8. CONTRACT OR GRANT NUMBER(s) N00014-67-0113-0029
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Psychology University of Minnesota Minneapolis, Minnesota 55455		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS P.E.:61153N PROJ.:RR042-04 T.A.:RR042-04-01 W.U.:NR150-343
11. CONTROLLING OFFICE NAME AND ADDRESS Personnel and Training Research Programs Office of Naval Research Arlington, Virginia 22217		12. REPORT DATE October 1975 13. NUMBER OF PAGES 45
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)

A conventional vocabulary test and two forms of a stradaptive vocabulary test were administered by a time-shared computer system to undergraduate college students. The two stradaptive tests differed in that one counted question mark responses (i.e., omitted items) as incorrect and the other ignored items responded to with question marks. Stradaptive test scores were more consistent with the hypothesized nature of the population distribution of verbal ability. When corrected for differing levels of item discrimination

3

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

and memory effects, the test-retest stabilities of the two testing strategies were about equal. Scores on one form of the stradaptive test were found to be very stable for testees who had highly consistent response records on initial testing. Stability of "subject characteristic curve" data was high, suggesting the usefulness of these data for describing test-testee interactions. Of the ten stradaptive ability scores studied, which grouped into four clusters, average difficulty scores had the highest stabilities. Analysis of difficulties of items associated with correct, incorrect, and question mark responses suggested that items with question mark responses should not be ignored, but should be treated as incorrect responses in branching decisions. Suggestions for future research on the stradaptive testing model are made.

CONTENTS

Introduction 1

Method 4
Design 4
Implementation of the Stradaptive Testing Strategy 4
Item structure 4
Scoring the stradaptive test 7
Ceiling and basal strata 7
Ability level scores 8
Consistency scores 10
A sample stradaptive response record 10
The Conventional Test 12
Subjects 12
Analyses 13
Comparison of the stradaptive test with the conventional test . 13
Descriptive statistics 13
Internal consistency 13
Test-retest stability: The problems 15
Test-retest stability: The corrections 16
Correlations between stradaptive and conventional tests .. 18
Further analyses of the stradaptive tests 18
Intercorrelations among stradaptive scores 18
Utility of the stradaptive consistency scores in predicting stability 18
Stability of stradaptive test response records 18
Relative difficulties of items producing different kinds of responses 19
Test length vs. ability 20

Results 20
Comparisons of the Stradaptive and Conventional Tests 20
Descriptive statistics 20
Internal consistency 22
Test-retest stability 23
Zero order stabilities 23
Partial correlations 26
Eta coefficients 26
Test-retest interval 26
Correlations between Stradaptive 1 and conventional test scores 26
Further Analyses of the Stradaptive Tests 28
Intercorrelations among scores 28
Utility of the stradaptive consistency scores in predicting stability 30
Stability of the stradaptive test response records 32
Stability of within strata data 32
Stability of total number administered 33
Relative difficulties of items producing different kinds of responses 34
Test length vs. ability 34

Summary and Conclusions 36

References 39

Appendix: Supplementary Tables 41

A STUDY OF COMPUTER-ADMINISTERED STRADAPTIVE ABILITY TESTING

In the early years of mental measurement, tests of individual differences were designed for individuals rather than groups. Binet's intelligence test, for example, was tailored to the individual. Using the Binet approach, the examiner neither wasted time administering items which were too easy for a testee nor frustrated the testee with items which were much too difficult. But the expense of the individual test administration required by Binet's approach forced test makers to devise an alternative measurement strategy that required less administration time by a trained psychometrist. The result was the group test.

In the process of applying group tests to the measurement of individuals, many of the advantages of individualized testing were sacrificed for the greater efficiency possible by measuring large groups of individuals at one time. The result was tests which were too difficult for some testees, and too easy for others, with measurement accuracy that varied widely as a function of ability level. Although group tests measured well the average member of the population for which they were constructed, there was still room for considerable improvement.

The advent of interactive computers provided an economical path for a return to individualized testing. With their development came the means to construct tests to efficiently measure individuals who were not necessarily typical of certain populations. A variety of techniques have been proposed for administering ability tests on interactive computer systems. Weiss and Betz (1973) summarize the recent literature on adaptive, or tailored, testing.

A basic premise of adaptive testing is that the best test for measuring an individual is a test with item difficulties peaked at the ability level of that individual rather than at the mean ability of a population. The fact that ability is not known until the end of the test has resulted in a diversity of strategies for choosing the items to be administered to a given individual (Weiss, 1974). These strategies can be divided into two subcategories--two-stage and multi-stage strategies. The latter are of the most interest here and can be further divided into two subclasses: variable branching procedures and fixed branching procedures.

Variable branching procedures include Bayesian and maximum likelihood approaches. Bayesian strategies, such as those proposed by Novick (1969) and Owen (1969), may begin with some initial estimate of the testee's ability, such as grade-point average or the testee's own subjective ability estimate. Given this ability estimate, every available item in the item pool is examined. Then, on the basis of the guidelines set by the particular model in use, the best item is chosen to be administered. Given the response to that item and the initial ability estimate, a new ability estimate is calculated and the procedure is repeated. The test usually terminates when a desired degree of precision of measurement is reached. Bayesian strategies have as their advantages the capability of using prior information about the ability of an individual, the tailoring of test length as well as difficulty, and, by examining all available items at each stage, the capacity to make very efficient

use of an item pool. Their disadvantage may lie in the failure of real items and real individuals to meet the crucial assumptions on which the Bayesian models are based. In addition, the computing time required to search a large item pool might lower their utility in interactive testing on small computer systems.

The fixed branching procedures use a set of items which are pre-structured by difficulty and/or discrimination. In these strategies, at any stage in the test a testee is branched to one predetermined category of items (which may consist of a single item) if he answers the item correctly, or to another predetermined category of items if he answers incorrectly. Because branching from any item is dependent only on the response to that item, the item pool does not need to be searched after each item response.

Most fixed branching procedures are variations of the pyramidal testing strategy (e.g., Larkin & Weiss, 1974). A pyramidal test has its items arranged in a triangular or pyramidal structure with item difficulties at the peak centered on the mean ability of the population of individuals to be measured (see Weiss, 1974, pp. 12-36). Difficulties increase or decrease with distance to the right or left of the peak. An individual taking a test under this strategy is first administered the item at the peak. If he answers it correctly, he is branched to the more difficult of the two items in the second stage; if he answers it incorrectly, he is branched to the less difficult item. This process continues until the testee reaches the end of a fixed number of stages.

Since each stage of a pyramidal test requires a number of items equal to the number of that stage, the pyramidal test requires a substantial number of items ($n(n-1)$ for an n-stage test). Furthermore, the pyramidal test is very inefficient in its use of available items. A pyramidal test has items at a number of difficulty levels. With the exception of the individual who answers all items correctly or all items incorrectly, a testee enters most difficulty levels somewhere after the first item at that difficulty level. Consequently, all the preceeding items at that difficulty level are not used. This is a problem in any real operationalization of the pyramidal strategy because there is no good position in the structure to put the most discriminating items. At no point in the structure will these items be routinely administered to testees whose sequence of item responses requires items to be administered at a given difficulty level.

Thus, the Bayesian strategies are promising because of their use of prior information, optimal branching, item economy, and flexible termination. But it remains to be seen whether the assumptions on which such strategies are based will be sufficiently met by real items and real individuals to realize an advantage in utility. The pyramidal strategy has as its advantage the lack of restrictive assumptions needed by the Bayesian strategies but lacks all the advantages of the Bayesian strategies--it makes no use of prior information, its termination criterion is inflexible, and it makes very inefficient use of an item pool. Clearly, some compromise approach is called for.

Such a strategy was proposed by Weiss (1973) and was named the stratified adaptive, or stradaptive, ability test. The stradaptive test is a collection of short peaked ability tests, each of these tests being referred to as a stratum. These strata are ordered by difficulty and are equally spaced along the ability continuum. Items within each stratum are ordered with the most discriminating items appearing first. Beginning with any rough ability estimate, a testee can begin the test in any stratum and is administered the most discriminating item in that stratum. On the basis of his response, the testee is branched either to a stratum with more difficult items, or to one with easier items, and is administered the most discriminating item in the chosen stratum. This process continues, with the testee being administered the most discriminating item yet unadministered in each stratum, until some termination criterion is reached. One termination criterion for the stradaptive test is based on a criterion borrowed from Binet. Its goal is to locate the level of chance responding, and termination occurs once this "ceiling level" is reliably located.

The stradaptive strategy bears some similarities to the Markov process with a reflecting barrier proposed by Mussio (1973), which was essentially a truncated pyramidal test. The stradaptive test is different in that it lacks Mussio's formal item structure, thus allowing better item economy, and lacks the common entry point and fixed number of items administered, which are characteristic of the pyramidal strategies.

The stradaptive test lacks the optimal branching of the Bayesian strategies but retains their advantages of untilization of prior information, tailored termination, and efficient use of the item pool. Its further advantage is that it does not require the restrictive assumptions on which the Bayesian strategies rest.

Waters (1974, 1975) reported the results of a study of live stradaptive ability testing. Using a pool of 250 verbal analogy items obtained from Educational Testing Service, he administered 46 conventional tests and 53 stradaptive tests to college students. His design allowed for the computation of both parallel forms reliability and validity coefficients. Validity was operationalized as a correlation between scores on his tests and scores on a conventional test composed of similar items taken earlier. His major findings were that 1) the stradaptive strategy was able to attain parallel forms reliabilities and validities comparable to a conventional test having twice as many items; 2) the reliability and validity of the stradaptive scores was strongly dependent on the termination criterion used; and 3) some methods of scoring the stradaptive test gave higher validities and reliabilities than other scoring methods, with the average difficulty of all items answered correctly consistently being one of the highest.

The present paper reports on the administration of two different stradaptive tests to college students to study the stradaptive strategy's psychometric characteristics, using an item pool and evaluative criteria different than those used by Waters. Further details on the logic and rationale of stradaptive testing are given in Weiss (1973).

METHOD

Design

This study was part of a larger research program studying the utility of computerized ability testing. One goal of the program is to determine the empirical relationships between ability estimates derived from the various adaptive testing strategies, as well as their relationships with ability estimates derived from a conventional test. In addition, strategies of adaptive testing are being evaluated in terms of other psychometric characteristics, in an attempt to identify those strategies which are most promising for practical applications. As part of that program, this study investigated the stradaptive testing strategy.

A 40-item conventional vocabulary test and two forms of a stradaptive vocabulary test were administered to college students. The two stradaptive forms differed in that one counted question mark responses (i.e., omitted items) as incorrect and the other ignored items responded to with question marks.

All tests were presented using Datapoint 3000 cathode-ray-terminals (CRTs) acoustically coupled to a Control Data Corporation 6400 time-shared computer. The testee responded on the CRT keyboard to each item presented with either a number indicating the multiple-choice alternative chosen or a question mark if he did not know the answer and chose not to guess. (See DeWitt and Weiss, 1974, for details of the test administration software.)

This study was concerned with two major kinds of analyses. First, data from the three tests were analyzed in terms of the characteristics of their score distributions, the correlations between stradaptive and conventional tests, and the magnitudes of their test-retest stabilities. Second characteristics of the stradaptive tests were investigated to provide a basis for refinement of the strategy. Among the characteristics investigated were the intercorrelations among the many methods of scoring a stradaptive test. This was done to determine which scores were redundant and could be eliminated. The utilities of the consistency scores in predicting test-retest stability were also investigated. To provide data for future development of subject characteristic curves (described below), stability of stradaptive test response records was investigated. The impact of ignoring omitted items was evaluated in terms of relative test-retest stabilities of scores derived from the two forms of stradaptive tests, and in terms of the relative difficulties of items giving rise to question mark responses. Finally, to evaluate the adequacy of the item pool (i.e., the effect of having many highly discriminating easy items but few highly discriminating difficult items), test scores were correlated with test length.

Implementation of the Stradaptive Testing Strategy

Item Structure

For this study, two forms of the stradaptive test were prepared and will

be referred to as Stradaptive 1 and Stradaptive 2. Stradaptive 1 is the stradaptive test used for illustrative purposes by Weiss (1973). It consisted of 229 vocabulary items taken from a larger pool of 369 items, with the restriction that items not overlap with those of a conventional test constructed for purposes of comparison. The larger pool was described by McBride and Weiss (1974), and norming item statistics for the 229 items used here are given in Appendix Table A-1.

Summary statistics for the items in both forms of the stradaptive test are given in Table 1. As is shown, the items in Stradaptive 1 were grouped into nine strata with stratum 5 centered on a normal ogive difficulty (Lord and Novick, 1968, pp. 376-378) of b=.007. The width of each stratum, and distance between the means of successive strata, was about 0.65 normal ogive difficulty units.

Stradaptive 2 consisted of 269 vocabulary items. This set of items was composed of most of the original 229 items of Stradaptive 1, the 40 items which were originally used in the conventional test, and a few new items. As can be seen from Table 1, the item structure of Stradaptive 2 was quite similar to that of Stradaptive 1, both consisting of items arranged in nine strata spaced about 0.65 difficulty units apart. Norming item statistics for Stradaptive 2 are also presented in Appendix Table A-1.

The most important distinction between the two stradaptive tests is the manner in which question mark responses (i.e., omitted items) were handled. In Stradaptive 1, a question mark was treated as an incorrect response. It caused the testee to be branched down one stratum, and was counted as incorrect when the scores were calculated. To investigate the effects of not penalizing the testee for answering honestly when he was not sure of the correct answer, question mark responses were ignored in Stradaptive 2. The subject was administered the next item in the same stratum (i.e., branched neither up nor down) and the item to which he responded with a question mark was not included in the calculation of scores.

The entry point or stratum in which the test was begun was determined for each testee using his reported grade-point average. The display presented on the CRT screen to the testee for this purpose, along with the entry stratum resulting from his response (which, of course, was not on the CRT screen) is shown in Figure 1.

Several branching rules were discussed by Weiss (1973) with respect to the stradaptive strategy and have been considered in discussions of other adaptive testing strategies (e.g., Weiss and Betz, 1973; Larkin and Weiss, 1974). The technique used here was the simple up-one, down-one branching rule. A testee was branched to the first unanswered item at the next more difficult stratum following a correct response, and to the first unanswered item at the next easier stratum following an incorrect response. The exception to this rule was when the testee gave a correct response to an item in the most difficult stratum or an incorrect response to an item in the least difficult stratum. In those instances, the testee was branched to the next item in the same stratum.

Table 1

Summary Statiatics for Conventional and Normal Ogive Item Parameters for Two Stradaptive Tests, by Stratum

Stratum	Stradaptive 1 No. of Items	Conventional p	Conventional r_{bis}	Normal Ogive b	Normal Ogive a	Stradaptive 2 No. of Items	Conventional p	Conventional r_{bis}	Normal Ogive b	Normal Ogive a
Stratum 1	35					35				
Mean		.949	.699	-2.648	1.290		.949	.699	-2.648	1.290
S.D.		.043	.210	.176	.925		.043	.210	.176	.925
High		.995	1.134	-2.393	3.000		.995	1.134	-2.393	3.000
Low		.850	.376	-2.980	.406		.850	.376	-2.980	.406
Stratum 2	36					33				
Mean		.863	.602	-1.926	.840		.863	.595	-1.951	.830
S.D.		.070	.161	.220	.383		.073	.166	.212	.398
High		.968	.869	-1.636	1.756		.968	.869	-1.657	1.756
Low		.709	.299	-2.322	.313		.709	.299	-2.322	.313
Stratum 3	36					39				
Mean		.763	.571	-1.287	.732		.771	.579	-1.314	.749
S.D.		.060	.130	.184	.266		.064	.129	.202	.263
High		.890	.813	-1.013	1.396		.890	.813	-1.013	1.396
Low		.628	.302	-1.627	.317		.628	.302	-1.653	.317
Stratum 4	30					46				
Mean		.631	.506	-.633	.648		.632	.499	-.666	.617
S.D.		.055	.116	.192	.284		.044	.098	.178	.236
High		.731	.680	-.343	1.822		.731	.680	-.343	1.822
Low		.542	.259	-.998	.268		.542	.288	-.998	.301
Stratum 5	25					44				
Mean		.498	.558	.007	.692		.503	.531	-.020	.643
S.D.		.043	.141	.189	.270		.042	.117	.196	.220
High		.568	.794	.329	1.306		.568	.794	.329	1.306
Low		.427	.331	-.285	.317		.427	.331	-.319	.317
Stratum 6	19					22				
Mean		.382	.469	.651	.549		.379	.455	.695	.527
S.D.		.038	.106	.185	.177		.036	.107	.206	.175
High		.434	.700	.977	.980		.434	.700	.977	.980
Low		.305	.346	.337	.369		.305	.296	.337	.310
Stratum 7	23					25				
Mean		.295	.436	1.327	.456		.296	.437	1.313	.460
S.D.		.039	.083	.183	.113		.039	.084	.182	.115
High		.353	.618	1.630	.718		.353	.618	1.630	.718
Low		.217	.323	1.004	.312		.217	.323	1.004	.312
Stratum 8	15					15				
Mean		.200	.427	2.006	.482		.200	.427	2.006	.482
S.D.		.047	.087	.206	.133		.047	.087	.206	.133
High		.274	.648	2.313	.851		.274	.648	2.313	.851
Low		.110	.321	1.649	.339		.110	.321	1.649	.339
Stratum 9	10					10				
Mean		.168	.387	2.621	.427		.168	.387	2.621	.427
S.D.		.069	.103	.273	.163		.069	.103	.273	.163
High		.300	.643	3.113	.840		.300	.643	3.113	.840
Low		.029	.253	2.320	.214		.029	.253	2.320	.214

Figure 1

Stradaptive Test Entry Point Question

IN WHICH CATEGORY IS YOUR CUMULATIVE GPA TO DATA?	Entry Stratum (not seen by student)
1. 3.76 to 4.00	9
2. 3.51 to 3.75	8
3. 3.26 to 3.50	7
4. 3.01 to 3.25	6
5. 2.76 to 3.00	5
6. 2.51 to 2.75	4
7. 2.26 to 2.50	3
8. 2.01 to 2.25	2
9. 2.00 OR LESS	1

ENTER THE CATEGORY (1 THROUGH 9) AND PRESS THE RETURN KEY.

Scoring the Stradaptive Test

Ceiling and basal strata. Several methods of scoring the stradaptive test require the use of ceiling and basal strata. These two concepts were borrowed from individual intelligence testing, primarily the Binet test. The basal level of responding is that difficulty level of items at which the testee answers all items correctly. The use of the basal level assumes that all less difficult items would also be answered correctly, and, therefore, easier items are not administered once a basal level has been established. The basal stratum was defined for use in the stradaptive test by Weiss (1973) as the most difficult stratum at which all items were answered correctly.

In the present data, if such a stratum existed it was identified as basal. If no stratum existed in which all items administered were answered correctly but at least one item was administered at the least difficult stratum, it was assumed that all strata were too difficult to be called basal and the hypothetical stratum below the lowest actual stratum was taken as basal. All other conditions (e.g., the response record was incomplete, there was no identifiable basal stratum, and no termination criterion had been reached) were considered abnormal terminations and the subject was eliminated. Most abnormal terminations were caused by computer failures, although a few were caused by subjects leaving early to meet other commitments.

The ceiling level of responding is that level of difficulty at which the testee answers no items correctly. This definition of the ceiling stratum assumes that he would answer no items correctly at any level of greater difficulty. Consequently, more difficult items are not administered. In the case of multiple-choice items, the testee is expected to answer some

items correctly due solely to chance successes. These chance successes are most likely to occur on items which are too difficult for a given testee. Thus, for multiple-choice items, the ceiling level can be defined as that level of difficulty where the testee answers correctly no more items than would be expected from random guessing.

In this study, which used five-alternative multiple-choice items, the ceiling stratum was defined as the least difficult stratum in which five or more items were administered and the testee answered 20% or less correctly. The five item minimum was established to allow a reasonably stable estimate of proportion correct at a given stratum for the critical termination criterion. If such a stratum existed, it was identified as the ceiling stratum. If no such stratum existed, but all items at the most difficult stratum had been administered, the hypothetical stratum immediately above the most difficult stratum was taken as the ceiling stratum. All other conditions were considered abnormal terminations and the testee was eliminated from all analyses.

Ability level scores. Weiss (1973) proposed ten methods of scoring the stradaptive test to obtain ability level estimates. These ability level scores are referred to by number in the figures and tables throughout this report. Score numbers and brief descriptions are shown together in the sample stradaptive test report shown in Figure 2.

Scores 1 through 3 are item difficulty scores. These scoring methods are borrowed from the pyramidal testing strategy (see Larkin and Weiss, 1974; Weiss, 1974, pp. 12-36). Score 1 is the difficulty of the most difficult item answered correctly. With the exception of abnormal terminations, this score could always be determined and was used as defined.

Score 2 is the difficulty of the (N+1)*th* item, or the next item that would have been administered had testing continued beyond termination. This score could not be determined in two circumstances. First, if termination was caused by running out of items in the next stratum to be drawn from, there obviously was no item from which to determine the score. Second, if the N*th* item was in the highest stratum and the response was correct, or the N*th* item was in the lowest stratum and the response was incorrect, the (N+1)*th* item would be chosen from a stratum that did not exist (i.e., a hypothetical stratum). In these cases, the effect would be the same as in the first situation (i.e., there would be no item from which to determine the score). In the first situation, where there was an insufficient number of items in an existing stratum, the average difficulty of the items in the stratum (the stratum difficulty) was substituted as the testee's score. In the second case, difficulties of hypothetical strata .65 units above the most difficult existing stratum or .65 units below the least difficult stratum were used as the testee's score.

Score 3 was defined as the most difficult non-chance item answered correctly. This was determined from the difficulty of the most difficult item answered in the stratum immediately below the testee's ceiling stratum. This item existed, and thus the score could be determined, except in the

condition where the ceiling stratum was stratum 1, the lowest actual stratum. In this case the difficulty of the lower hypothetical stratum was used.

Scores 4 through 6 can be referred to as stratum scores. These are stratum difficulty analogues to the three item difficulty scores. Score 4 was defined as the mean difficulty of the items at the most difficult stratum in which at least one item was answered correctly. Score 5 was defined as the mean difficulty of the stratum containing the $(N+1)th$ item (or hypothetical item if no item existed). Score 6 is the mean difficulty of the highest non-chance stratum or the stratum immediately below the ceiling stratum. These three scores, barring abnormal termination, were always determinable and were implemented as defined.

Score 7, the interpolated stratum difficulty score, was an attempt to determine the exact stratum difficulty at which the testee would respond at a chance level when that difficulty fell between two available stratum difficulties. Algebraically, it was defined as:

$$A=\overline{D}_{c-1} + S\ (P_{c-1} - .50)$$

where: $\overline{D}_{c-1}$ is the average difficulty of the $(c-1)th$ stratum and c is the ceiling stratum. It is therefore, the average difficulty of all items available at the testee's highest non-chance stratum, or the stratum just below his ceiling stratum.

P_{c-1} is the testee's proportion correct at the $(c-1)th$ stratum and S is $\overline{D}_c - \overline{D}_{c-1}$, if P_{c-1} is greater than .50, or $\overline{D}_{c-1} - \overline{D}_{c-2}$ if P_{c-1} is less than .50, where $\overline{D}$ is the average difficulty of the designated stratum.

It was possible to calculate this score except in the condition where the ceiling stratum was stratum 1. In that case, no proportion correct was available for the "c-1" stratum and the score could not be calculated. In this study, this particular condition never occurred. Thus, with the exception of abnormal terminations, score 7 was determinable for all testees.

Finally, three average difficulty scores were defined. Score 8 was defined as the average difficulty of all items answered correctly and was calculable in all cases. Score 9 was defined as the average difficulty of items correct between, but not including, the ceiling and basal strata. The hypothesized advantage of this score over score 8 was that it would be less susceptible to bias from inappropriate entry points. This score could be determined except when no items were answered correctly between the ceiling and basal strata, a condition caused by the ceiling and basal strata being adjacent. When this occurred, score 9 was not calculated. Score 10 is the average difficulty of all items answered correctly at the

highest non-chance stratum and was calculable except when the ceiling stratum was stratum 1, a condition not encountered in this study.

Consistency scores. Weiss (1973) suggested that the consistency of the response record, or variability of difficulties of items encountered by a given testee, might yield information about the confidence which could be placed on the point ability estimates obtained from the first ten scores. Specifically, he said (p. 26), "Individuals who are more consistent should have more stable ability estimates, while those who are less consistent should have less stable ability estimates." This hypothesis was studied using five scores designed to reflect response consistency.

Two consistency indices reflect the overall variability of the difficulties of the items administered to a given testee. Score 11 is defined as the standard deviation of the difficulties of all items administered. Except for abnormal terminations, this score was always calculable. Score 12 is defined as the standard deviation of item difficulties of all items answered correctly. In this study this score also was available for all testees.

In an attempt to control for inappropriate entry points, three indices reflect consistency using an individual's ceiling and basal strata. Score 13 is defined as the standard deviation of difficulties of all items answered correctly between the ceiling and basal strata. This score could not be calculated for a given testee when less than two items were answered correctly between the ceiling and basal strata, a condition always caused by the ceiling and basal strata being adjacent. Score 14 is defined as the difference in average stratum difficulties of the ceiling and basal strata. Scores 14 and 15 have an advantage over score 13 in that they are calculable for all testees.

A Sample Stradaptive Response Record

Figure 2 shows the stradaptive test performance of a college sophomore. This test record is typical of the stradaptive test performance of college students. The testee was first presented with an entry point screen (Figure 1) and indicated that his cumulative grade-point average to date was between 2.76 and 3.00. He thus began the stradaptive test at stratum 5. His answer to the first item was correct (indicated by a "+" in Figure 2), which branched him to the first available item in stratum 6. Correct answers to the second and third items resulted in his moving to stratum 8, where he received the first item from that more difficult peaked test. Since the stage 4 item was too difficult for him, his response was incorrect (-), and he branched downward to the second item in stratum 7. The student then alternated between correct and incorrect responses for the items at stages 6 through 8, followed by an incorrect response to the stage 9 item. This returned him to stratum 6 for his tenth item. With a few minor deviations, he then essentially alternated between correct and incorrect responses from stages 11 through 20. Item 20 terminated the stradaptive test since the testing procedure had, at that point, located the student's ceiling stratum; at stratum 8 he had answered incorrectly all five items.

Figure 2

Report on a Stradaptive Test for a Consistent Testee

REPØRT ØN STRADAPTIVE TEST

ID NUMBER:

DATE TESTED: 73/07/12

STRATUM:	(EASY) 1	2	3	4	5	6	7	8	(DIFFICULT) 9
	1+	2+	3+	4-					
			5+	6-					
			7+	8-					
			9-						
		10+	11+	12-					
			13-						
		14+	15-						
		16+	17-						
		18+	19+	20-					
PRØP.CØRR:					1.00	1.00	.56	0.00	

TØTAL PRØPØRTIØN CØRRECT= .550

SCØRES ØN STRADAPTIVE TEST

Ability Level Scores

1. DIFFICULTY ØF MØST DIFFICULT ITEM CØRRECT= 1.49
2. DIFFICULTY ØF THE N+1 TH ITEM= 1.44
3. DIFFICULTY ØF HIGHEST NØN-CHANCE ITEM CØRRECT= 1.49
4. DIFFICULTY ØF HIGHEST STRATUM WITH A CØRRECT ANSWER= 1.33
5. DIFFICULTY ØF THE N+1 TH STRATUM= 1.33
6. DIFFICULTY ØF HIGHEST NØN-CHANCE STRATUM= 1.33
7. INTERPØLATED STRATUM DIFFICULTY= 1.37
8. MEAN DIFFICULTY ØF ALL CØRRECT ITEMS= .88
9. MEAN DIFFICULTY ØF CØRRECT ITEMS BETWEEN CEILING AND BASAL STRATA= 1.26
10. MEAN DIFFICULTY ØF ITEMS CØRRECT AT HIGHEST NØN-CHANCE STRATUM= 1.28

Consistency Scores

11. SD ØF ITEM DIFFICULTIES ENCØUNTERED= .59
12. SD ØF DIFFICULTIES ØF ITEMS ANSWERED CØRRECTLY= .46
13. SD ØF DIFFICULTIES ØF ITEMS ANSWERED CØRRECTLY BETWEEN CEILING AND BASAL STRATA= .18
14. DIFFERENCE IN DIFFICULTIES BETWEEN CEILING AND BASAL STRATA= 1.36
15. NUMBER ØF STRATA BETWEEN CEILING AND BASAL STRATA= 1

The Conventional Test

As in previous studies in this series (Betz and Weiss, 1973, 1974; Larkin and Weiss, 1974), a conventional test was administered by computer to provide a comparison with the stradaptive tests. This was a 40-item peaked test for which norming summary statistics are shown in Table 2. Item

Table 2

Summary Statistics for Item Parameters of the 40-Item Conventional Test

Parameters	Mean	S.D.	High	Low
Traditional				
p, proportion correct	.537	.010	.661	.267
bis, biserial *r* with total score	.472	.078	.612	.296
Normal Ogive[a]				
b, difficulty	-.188	.592	1.155	-.956
a, discrimination	.543	.112	.774	.310

[a] Estimated using formulas described by McBride and Weiss (1974, p. 24)

statistics for this test are presented in Appendix Table A-2. According to Betz and Weiss (1973) who constructed the test, "Items were selected from the pool [of 369 items] that had difficulties closest to p=.55 and item total score biserial correlation coefficients closest to .45," (p. 15). The score for the conventional test was the proportion of items answered correctly by each testee.

Subjects

Subjects providing the data for this study were college students. Some sophomores were recruited from the psychology department's subject pool, but the majority of the testees were juniors, seniors, and first-year graduate students from courses in psychological statistics and measurement. To obtain test-retest stability data, some subjects were tested and then retested after an interval of from two to 11 weeks. Valid test-retest data were collected on 180 testees for Stradaptive 1, 98 testees using Stradaptive 2, and 194 testees using the conventional test. To obtain the best possible distributional and intercorrelational data on the stradaptive tests, single administration data were gathered from other facets of the research program's general data collection yielding initial test data on 476 students for

Stradaptive 1 and 113 students for Stradaptive 2.

Analyses

Comparison of the Stradaptive Test with the Conventional Test

These analyses were designed to investigate the characteristics of the various scores derived from the stradaptive test, in comparison to scores derived from the conventional test. Of interest were the characteristics of their score distributions, as well as inter-relationships among scores derived from the two testing strategies. In addition, the relative stability of ability estimates derived from the two testing strategies was considered important, as it was in previous studies (e.g., Betz and Weiss, 1974; Larkin and Weiss, 1974). Score stability was viewed both as an indication of the relative reliabilities of ability estimates derived from the two testing strategies and as an indication of the practical utility of the ability estimates for making longitudinal predictions.

Descriptive statistics. The mean, standard deviation, skewness, and kurtosis were calculated for all score distributions obtained from Stradaptive 1, Stradaptive 2, and the conventional test. The underlying ability distribution of the population sampled was not known. Therefore, scores derived from the different testing strategies could not be evaluated on the basis of how well they reflected the distribution of true ability. A normally distributed score distribution is statistically convenient, however, and for this reason alone, scores that depart radically from normality are undesireable.

Internal consistency. In practical testing applications, internal consistency is calculated as an approximation to parallel forms reliability. Ability tests are typically constructed to maximize internal consistency. In inter-strategy comparison research, such as is reported here, the goal is to equate internal consistency across strategies. Given a unidimensional trait, the internal consistency is partly a function of the discriminating power of the items (Gulliksen, 1950). Thus, testing strategies which are equated for internal consistencies can then be meaningfully compared in terms of stabilities, since all strategies will have equally good items.

Internal consistencies were calculated in this study for both the conventional and adaptive tests. These data were then used as a basis for statistically equating the discriminating power of the item pools to provide a more realistic comparison of the test-retest stabilities of the two testing strategies.

Calculation of the internal consistency reliability of an adaptive test cannot use standard approaches because 1) all individuals do not encounter the same items, and 2) those items they do encounter cannot be thought of as a random sample from the total pool. Some adaptive testing strategies, such as two-stage, allow internal consistency to be calculated on subgroups of items (e.g., Betz and Weiss, 1973) but this is usually an underestimate due to restriction of range in ability. Larkin and Weiss (1974) were able to calculate internal consistency for a pyramidal test using a scoring technique that

predicted a testee's response to all those items of the test which he had not actually encountered. They concluded, however, that the resulting internal consistency was an overestimate due to the assumptions made by the scoring technique.

A different approach to estimating the internal consistency of an adaptive test was taken in this study. Gulliksen (1950) presents a formula for calculating internal consistency reliability from item reliabilities (i.e., the item-total correlation weighted by the item variance). The formula (Equation 21, p. 378) is a variation of the Kuder-Richardson formula 20 (KR-20):

$$r_{xx} = \frac{k}{k-1}\left[1 - \frac{\sum_{g=1}^{k} s_g^2}{\left(\sum_{g=1}^{k} r_{xg}\, s_g\right)^2}\right] \qquad [1]$$

where r_{xx} ≡ internal consistency of the test

k ≡ number of items in the test

s_g^2 ≡ item variance $= p(1-p)$, where p = proportion correct

and r_{xg} ≡ correlation of item response and total score

This formula, as derived by Gulliksen, is strictly correct mathematically only when it is used to calculate test reliability from the same sample of items on which the item reliabilities were calculated, and hence offers no direct advantage over the KR-20. But a reliability coefficient can be obtained by assuming that item-total correlations and proportion correct data obtained in the norming study (McBride and Weiss, 1974) are acceptable estimates of the item-total correlations that would have been obtained if a representative sample of the population of individuals had been given the items of interest. As a further departure from standard usage of this formula, the biserial rather than the point-biserial item-total correlation was used for the calculations. Although the formula was derived in terms of point-biserial correlations, the point-biserial is affected by the difficulty of the items, dropping when items are very easy or very hard. Thus, the biserial correlation is more appropriate for use in an adaptive test, since item difficulties (and, therefore, item-total point-biserial correlations) will vary with ability levels.

In this study, the internal consistency coefficient was calculated for the stradaptive tests as follows: Substituting norm group item parameters into Gulliksen's formula, a reliability coefficient was calculated for each person's set of items. This reliability was then inflated or deflated to a length of 29 items (the mean length of the stradaptive test) using the Spearman-Brown formula. These coefficients were then averaged across all individuals

using an r to z transformation. This yielded an internal consistency coefficient characteristic of a set of 29-item conventional tests assembled from the stradaptive item pool with test difficulties distributed as a function of the underlying ability.

To determine the utility of this technique, it was tested empirically. Internal consistency was calculated by Equation 1 for four subsets of 10, 20, 30 and 40 items from the conventional test. The coefficients obtained were inflated to a length of 40 items and compared with each other and with a Hoyt internal consistency reliability coefficient calculated on the total conventional test.

Test-retest stability: the problems. Test-retest stability coefficients were of prime interest in this study as a means of comparing the relative precision and practical utility of scores resulting from the stradaptive and conventional testing strategies. Unfortunately, the conventional test was constructed to match the psychometric characteristics of a two-stage test and its match with the stradaptive tests was something less than optimal. The first problem encountered with the conventional test was the fact that it was longer than the typical stradaptive test. The average length of Stradaptive 1 was 27.75 items on initial testing and 31.35 items on retest. Average lengths for Stradaptive 2 were 25.38 and 26.61 when question mark responses were not counted. When question mark responses were counted, those lengths rose to 29.23 and 30.64. The 40-item conventional test had the clear advantage with respect to test length.

Also in favor of the conventional test, with regard to estimating test-retest reliabilities, was the fact that it had all forty initial test items repeated on retest, thus inflating the test-retest correlation because of memory effects. The existence of memory effects with these items was demonstrated by Betz and Weiss (1973) and Larkin and Weiss (1974).

Working against the conventional test was the fact that its item discriminations were lower than those of the stradaptive tests. The average normal ogive discrimination for conventional test items was a=.543. The average discrimination for all the items in the stradaptive pools were a=.746 and a=.717 for forms 1 and 2 respectively. However, since the stradaptive item pool was constructed so that the most discriminating items are administered first, the average discrimination of the items actually administered was higher than the average discrimination of all items in the item pool. The average discriminations of all items administered, each item weighted by the number of times it was administered, were a=.841 and a=.879 for forms 1 and 2 respectively. This result clearly favored the stradaptive tests.

The final inequity was that the stability of the stradaptive tests was influenced to some degree by the use of initial ability estimates for entry points. The initial ability estimate obtained prior to the first testing was used on both the initial test and the retest. Therefore, as the test length approached zero items, the stability approached unity. Although the shortest test contained nine items, this factor still likely had some influence on the test-retest stability of stradaptive scores.

Test-retest stability: the corrections. A length of 29 items was taken as an approximate average of the lengths of the stradaptive tests. This is eleven items shorter than the forty-item conventional test.

While the conventional test had all its items repeated on retest, the stradaptive tests rarely had the initial item set repeated on retest. To calculate the proportion of items encountered in both initial test and retest for Stradaptive 1, the smaller number of items within a stratum, on test or retest, was taken as the number of common items in that stratum. This number was summed over all strata and all individuals, and was divided by one half the total number of items administered on test and retest to all individuals who took the Stradaptive 1 test. The proportion of common items for Stradaptive 2 was calculated in the same way. But where the Stradaptive 1 calculation gave an exact figure for number of items encountered twice, the Stradaptive 2 calculation yielded only an approximation. This was because totals within strata for Stradaptive 2 did not include question mark responses. The proportion of items common on test and retest was .615 for Stradaptive 1 and .567 for Stradaptive 2.

Within the conventional test, memory effects and lengths were equated simultaneously by preparing, from the original set of forty items, five analogous test pairs. Each pair consisted of one randomly selected test of twenty-nine items and a second test containing the remaining eleven items and eighteen of the first tests' twenty-nine items. This yielded five pairs of twenty-nine item tests, each pair having 18 or 62% of their items in common, thus matching the average proportion of items in common on the Stradaptive 1 retest. Items for one test in each pair were scored from the initial test data and items for the other were scored from the retest data. As an estimate of stability of such an analogous form, the mean (r to z transformed) correlation between members of the five pairs was used.

A direct correction for the effects of differences in item discrimination on test-retest reliability was not available. A correction was implemented, however, based on the fact that item discrimination has an effect on internal consistency reliability, which has an effect on validity. It was further assumed that correlational validity is in some respects analogous to test-retest reliability. Gulliksen (1950) provided a formula (eq. 8-19, p. 83) for calculating the necessary increase in test length to obtain a desired internal consistency:

$$K = \frac{(1-r)R}{(1-R)r} \qquad [2]$$

where $K \equiv$ proportionate increase in length
$r \equiv$ the original internal consistency
$R \equiv$ the desired internal consistency

He also provided a formula (eq. 9-19, p. 98) to predict the change in validity of one test in predicting another as a function of changes in the lengths of both tests. In the case of stability coefficients where both tests are the same and both lengthened the same amount, that formula becomes:

$$r'_{tt} = \frac{r_{tt}}{\frac{1}{K} + \left(1 - \frac{1}{K}\right) r_{xx}} \qquad [3]$$

where r'_{tt} ≡ corrected test-retest correlation

r_{tt} ≡ original test-retest correlation

r_{xx} ≡ original internal consistency

K ≡ proportionate increase in test length from previous equation

Equation 3 may be recognized as a variation of the Spearman-Brown formula.

To correct for unequal discriminations, the conventional test internal consistency calculated using the norming parameter method described earlier was substituted for the original internal consistency in Equation 3. The average internal consistency of the stradaptive tests, the calculation of which was described earlier, was substituted for the desired internal consistency. From this, the proportionate increase in length of conventional test required to compensate for different discriminations was calculated. Then the average stability of the five pairs of analogous conventional tests was inflated using Equation 3 to the value expected had either the tests been lengthened to compensate or the discriminations been equivalent.

It should be noted that the presence of these many corrections precludes the drawing of any strong conclusions from this study regarding test-retest stability. Several stability coefficients were calculated, however. Both test-retest product-moment correlations and eta coefficients were calculated for the forty-item conventional test and the five pairs of analogous forms.

Finally, to assess the maximum inflation of the stradaptive stability coefficient that could be caused by the initial ability estimates, partial correlations between test and retest administrations of the two stradaptive tests were calculated, with initial ability estimate partialled. The partial correlation is probably an underestimate of the stability coefficient that would have been obtained had initial ability estimates actually been held constant. The reason for this is that the initial ability estimate has both valid and error variance associated with it, and both the valid as well as the error variance are removed by the partialling procedure. This partialling problem was discussed in detail by Meehl (1970). For purposes of comparison, the initial ability estimates were also partialled out of the conventional test stabilities. The correlation between conventional test score and initial ability estimate can be construed as common variance due to the underlying ability, and any reduction in conventional test-retest correlation reflects how much the stradaptive reliabilities were artifactually deflated in the partialling process.

Both to make the stabilities more comparable and to observe the effect of time on stability, testees were divided into subgroups according to the length of the test-retest interval: 0-15, 16-30, 31-45, 46-60 and over 60 days. Product-moment stability coefficients were then calculated using

scores of those testees within each time group from both Stradaptive 1 and conventional tests. Stradaptive 2 data were not included in this analysis because virtually all test-retest intervals fell into one of the above groups (thus precluding trend analysis) and too few to analyze meaningfully fell into a time period overlapping with a period from the other two tests (thus precluding analysis within comparable intervals).

Correlations between stradaptive and conventional tests. Stradaptive 1 scores were correlated with the forty item conventional test scores for those testees who completed both on the same occasion. This correlation was computed to determine whether the stradaptive and conventional tests were measuring the same ability. Stradaptive 2 scores were not correlated with the conventional test score because no subjects were given both the Stradaptive 2 test and the conventional test.

Further Analyses of the Stradaptive Tests

Intercorrelations among stradaptive scores. Intercorrelations among scores on the stradaptive test were calculated for the initial administrations of both stradaptive tests. This was done to provide a basis for reducing the number of scoring methods. If several scores are to be calculated, they must be sufficiently independent of each other in order to provide differential information.

Utility of the stradaptive consistency scores in predicting stability. The five consistency scores were proposed as predictors of stability of the ability scores. To determine whether a consistency score functioned in this manner, subjects were first divided into five groups on the basis of that score on initial testing, and then within-group stability analyses were performed. Specifically, Stradaptive 2 testees were first ranked on the basis of a consistency score. This distribution was then divided into five groups with approximately equal numbers of testees. Stradaptive 1 testees were then grouped on the basis of cutting scores established in the Stradaptive 2 division. Stradaptive 2 was chosen for the initial division in order to provide a sufficient number of subjects in each group to allow meaningful analysis, since the total number of subjects who completed Stradaptive 2 was smaller. After division into sub-groups, product-moment test-retest coefficients were calculated within each of these groups, ranging from a group of highly consistent testees to a group of highly inconsistent testees.

This analysis was performed on only three of the consistency scores--scores 11, 12, and 13. The scores analyzed were chosen because they were all standard deviation scores and this allowed a direct comparison of scores 11 and 12, the overall variability scores, with score 13, a statistically similar score based on variability between ceiling and basal strata.

Stability of stradaptive test response records. Weiss (1973) suggested that ability scores might be estimated from a testee's stradaptive test response record using "subject characteristic curves". These curves are analogous to "trace lines" and are based on a testee's obtained proportion

correct at each stratum. He suggested that analysis of these data to obtain ability estimates might proceed along the lines of estimating normal ogive item parameters. Such latent parameters were not estimated in this study. But, to facilitate future research into the utility of such data, several indices of the stability of the stradaptive test response record were computed. Thus, the stability of stradaptive test length was determined from a product-moment correlation between number of items administered on initial test and on retest.

No common index existed for overall stability of the subject characteristic curve data, as reflected in total number of items answered within strata or proportions correct within strata. The form of the data, however (multiple continuous predictors and criteria), suggested the canonical correlation model. Thus, canonical correlations between test and retest data were computed. Two canonical analyses were implemented, using as variables in one analysis the number of items administered in each stratum and, in the second analysis, the proportions correct at each stratum.

For this canonical analysis, there were usually several strata in which no items were administered and thus proportions correct could not be calculated. To remedy this, the proportions correct below the ceiling stratum were set to zero. Zero was used rather than the chance level because, in the stradaptive testing strategy using an up-one, down-one branching strategy, unless the testee runs out of difficult strata, he gets no items correct at his highest stratum.

A complete redundancy analysis (Stewart and Love, 1968; Weiss, 1972) was performed on the canonical correlations. The redundancy index of greatest interest here is the redundancy of the retest given the initial test. This can be interpreted as the proportion of variance in the retest data predictable from the initial test data. It is also interpretable as the average squared multiple correlation of scores on each retest stratum with all scores on initial test strata. This redundancy coefficient bears some similarity to a test-retest reliability coefficient, but it expresses the stability of characteristics of the response records on the stradaptive test rather than merely the stability of summary scores as does the traditional test-retest reliability coefficient.

Relative difficulties of items producing different kinds of responses. One objective of this study was to examine the effects of not penalizing testees for honestly admitting they were not sure which multiple-choice answer was correct. This comparison was possible since Stradaptive 1 was designed to treat a "?" response as incorrect (thereby branching to a less difficult item) while Stradaptive 2 treated the same response as "no information" and presented another item at the same stratum.

In addition to the test-retest data showing the relative stabilities of scores on the two forms of the stradaptive test, an analysis was done to determine if the average difficulty of items answered with a question mark was equal to a testee's ability, or more difficult than his ability, and if more difficult, how much more difficult.

Score 8, the average difficulty of all items correct, was used as an estimate of ability. The difficulty of each item administered to an individual was deviated from that individual's ability level (operationalized as Score 8). These deviated difficulties were grouped into difficulties of correct, incorrect, and question mark response items and then pooled over all individuals for Stradaptive 1 and Stradaptive 2 administrations separately. Both initial test and retest data were used. Means of these deviated difficulties yielded the average distance from ability, in normal ogive difficulty units, of items generating the various types of responses. Standard deviations of the deviated difficulties were also computed.

Test length vs. ability. Ability scores derived from stradaptive testing were correlated with test length. This analysis was designed to determine whether there were interactions of scoring methods with characteristics of the item pool which resulted in different correlations of scores derived from each method with the number of items required to reach the termination criterion. A slight correlation was expected because the more discriminating items available at the lower difficulty strata were expected to yield fewer incorrect branchings and thus faster terminations.

RESULTS

Comparison of the Stradaptive and Conventional Tests

Descriptive Statistics

Descriptive statistics for the initial testing of Stradaptive 1, Stradaptive 2, and the conventional tests are shown in Table 3. Retest data are summarized in Appendix Table A-2. Standard deviations of ability scores and of consistency scores 11 to 14 are in normal ogive scoring units and are thus comparable. Score 15 is in stratum units but its standard deviation was multiplied by .65 (the width of a stratum in normal ogive difficulty units) to be comparable to the other scoring methods.

The ten Stradaptive 1 ability scores show roughly equal standard deviations. Most of these ability scores had a significant positive skew and all were significantly platykurtic. The distance consistency scores, scores 14 and 15, show higher means and larger standard deviations than the standard deviation consistency scores, scores 11, 12 and 13. All consistency scores were significantly positively skewed. The overall variability consistency scores, scores 11 and 12, were leptokurtic, and the between ceiling and basal indices, scores 13, 14, and 15, ranged from normal to significantly platykurtic.

Means of Stradaptive 2 ability scores were consistently lower than Stradaptive 1 scores. Standard deviations were also consistently smaller. All ability scores were positively skewed, and although the values were higher than for Stradaptive 1, fewer were significant due to the smaller number of subjects. None of the ability scores distributions deviated significantly from normality in terms of kurtosis. Characteristics of the consistency score distributions were similar for both forms of the Stradaptive test.

Table 3

Characteristics of Score Distributions for Stradaptive 1, Stradaptive 2, and the Conventional Test on Initial Testing

Score	N	Mean	S.D.	Skew	Kurtosis
Stradaptive 1					
Ability Scores					
1	476	1.073	1.187	-.080	-.785*
2	476	.560	1.468	.402*	-.714*
3	476	.531	1.324	.313*	-.680*
4	476	1.019	1.148	-.306*	-.893*
5	476	.570	1.453	.350*	-.680*
6	476	.370	1.274	.172	-.797*
7	476	.440	1.241	.229*	-.757*
8	476	-.042	1.055	.324*	-.613*
9	420	.066	1.122	.340*	-.507*
10	475	.339	1.270	.209	-.770*
Consistency Scores					
11	476	.753	.186	.661*	.780*
12	476	.661	.211	.570*	1.322*
13	420	.380	.219	.348*	-.605*
14	476	1.925	.857	.569*	-.034
15	476	1.947	.847	.571*	-.007
Stradaptive 2					
Ability Scores					
1	113	.774	1.064	.305	-.527
2	113	.173	1.212	.537*	.571
3	112	.167	1.084	.636*	-.048
4	113	.748	1.047	.176	-.871
5	113	.188	1.179	.579*	.790
6	113	-.006	1.120	.355	.170
7	113	.085	1.077	.445	.233
8	113	-.350	.853	.442	-.376
9	94	-.241	.944	.622*	.126
10	112	-.004	1.077	.566*	-.117
Consistency Scores					
11	113	.752	.196	.978*	1.524*
12	113	.667	.225	.406	.785
13	94	.389	.195	.274	-.491
14	113	1.815	.832[a]	.523*	.113
15	113	1.788	.822[a]	.538*	.171
Conventional Test					
40-items	194	.588	.209	-.110	-.945*
29-item analogous form[b]	194	.588	.213	-.129	-.884*

[a] S.D. is multiplied by .65.

[b] All statistics for the 29-item analogous form are means of statistics calculated on five combinations of items.

*Significantly different from zero at $p<.05$.

Score distributions were essentially the same for the forty-item conventional test and the twenty-nine item analogous forms. Both distributions were symmetric around means of .588 and both were significantly flat.

Because of differences in scoring methods, no direct comparisons of means and standard deviations can be made between the conventional and stradaptive tests. The noticeable positive skew of the stradaptive scores was absent in the conventional test data. Platykurtosis was similar in Stradaptive 1 and the conventional test, but was not evident in Stradaptive 2.

Internal Consistency

Table 4 presents the internal consistency of the conventional test calculated using the item reliability formula, Equation 1 on page 11, from subsets of 10, 20, 30 and 40 items. Also shown is the internal consistency calculated from Hoyt's formula using all 40 items. Since the average length of the Stradaptive test was about 29 items, the estimated reliability for a conventional test of 29 items (shown in row 2 of Table 4) is most relevant for this study. No definite trend is apparent in the corrected coefficients as the number of items used is increased from 10 to 40. The coefficients do appear to be higher than those obtained from Hoyt's method, however.

Table 4

Internal Consistency of the Conventional Test as Estimated from Subsets of Items Using the Item Reliability Method, Compared with Internal Consistency Calculated Using Hoyt's Method

	Number of Items				
Internal Consistency	10	20	30	40	Hoyt[a]
Of item samples	.686	.833	.887	.911	.893
Corrected to 29 item length	.863	.879	.883	.881	.858
Corrected to 40 item length	.897	.909	.912	.911	.893

[a]Based on 40 items.

Internal consistencies of the stradaptive tests calculated from the item reliability formula are presented in Table 5. As can be seen, they are

substantially higher than those of the conventional test (as shown in Table 4), a result due to the higher item discriminations in the stradaptive item pools.

Table 5

Average Internal Consistencies of the Stradaptive Tests (Unweighted r to z Averages of Internal Consistencies)

Corrected Length	Stradaptive 1	Stradaptive 2	Average
29 items	.935	.942	.938
40 items	.952	.957	.954

Test-retest Stability

Test-retest stability coefficients for the stradaptive tests are presented in Table 6. Conventional test coefficients are presented in Table 7. These tables contain zero-order product-moment correlations between test and retest, partial correlations between test and retest with initial ability estimates held constant, and eta coefficients.

Zero-order stabilities. Based on the zero-order correlations shown in Table 6, the average difficulty scores, scores 8 and 9, were the most stable ability scores on both forms of the stradaptive test. Scores 2 and 5, the (N+1)*th* item and stratum scores, were the least stable scores on both forms. The remainder of the ability scores fell between these.

These differences between stabilities of the stradaptive ability scores appear to be a function of the amount of information used by the scores. The average difficulty scores are highest because they make use of information gained from all items administered. The (N+1)*th* item and stratum scores, on the other hand, are least stable because they are heavily dependent on the response to the last item. A correct response on the final item makes these scores two stratum units higher (1.30 normal ogive difficulty units, or 20% of the score range, based on score ranges of -3.25 to 3.25) than does an incorrect response to the same item.

As Table 6 shows, stabilities of the consistency scores were much lower than those of the ability scores. For the Stradaptive 1 data, the stabilities of the overall variability scores, scores 11 and 12, were highest at .569 and .496 respectively. The stabilities of the scores representing variability between

Table 6

Test-Retest Stabilities of Stradaptive Tests

Stradaptive 1			
Ability Scores	Zero-Order Correlation (N=170)	Partial Correlation (N=167)	Eta Coefficient (N=170)
1	.838	.814	.838
2	.787	.774	.805
3	.845	.833	.849
4	.833	.810	.833
5	.787	.773	.787
6	.841	.829	.841
7	.872	.861	.878
8	.920	.901	.920
9	.912	.902	.912
10	.842	.829	.851
Consistency Scores			
11	.569	.577	.577
12	.496	.485	.569
13	.252	.234	.327
14	.328	.324	.398
15	.321	.318	.364

Stradaptive 2			
Ability Scores	Zero-Order Correlation (N=79)	Partial Correlation (N=76)	Eta Coefficient (N=79)
1	.717	.678	.766*
2	.630	.601	.647
3	.741	.733	.741
4	.690	.642	.734*
5	.654	.625	.695*
6	.717	.708	.738
7	.741	.730	.778
8	.823	.789	.823
9	.811	.792	.811
10	.746	.737	.746
Consistency Scores			
11	.510	.510	.553
12	.300	.298	.385
13	.110	.134	.337
14	.128	.171	.274
15	.120	.164	.152

Note: N's reported refer to the number of subjects with valid data. Stabilities for some scoring methods are based on fewer subjects.

*Curvilinearity significant at $p<.05$

ceiling and basal strata, socres 13, 14 and 15, were too low to consider those scores representative of a stable trait. This does not necessarily imply, however, that the consistency scores will not be useful in other ways, such as moderating test-retest reliability of the ability scores.

Three zero-order correlations are shown in Table 7 for the total group who completed the conventional test: 1) the test-retest coefficient for the entire 40-item test; 2) the average test-retest coefficient for the five analogous forms (correlations for the five pairs were .876, .873, .871, .862, and .890); and 3) the average correlation corrected to equate internal consistencies of the stradaptive and conventional tests. The

Table 7

Test-Retest Stabilities of the Conventional Test

Form	Total Group (N=194)	Testees with Initial Ability Estimates Available (N=81)	
		Zero-Order Correlation	Partial Correlation
40-item	.913	.900	.889
29-item analogous form	.874	.868	.856
29-item analogous form corrected for internal consistency	.931	...	...

corrected value was obtained by inflating the test-retest correlation using the method described above, from its value given a conventional test with internal consistency of .881 (the internal consistency of a 29-item test) to the value expected with a conventional test having an internal consistency of .938 (the average internal consistency of the stradaptive tests). The first two coefficients, while statistically more sound, are psychometrically inadequate for comparison. The latter value, while theoretically the most adequate of the three for comparison with the stability of the stradaptive test, rests on many corrections and assumptions. It is, therefore, only a rough estimate of the stability of a conventional test psychometrically equivalent to the stradaptive tests except for strategy of administration. As can be seen from Tables 6 and 7, the corrected stability of the analogous form of the conventional test (r=.931) is slightly higher than the highest stradaptive ability score stability (r=.920).

Partial correlations. Even though partial correlation analysis removed valid as well as extraneous variance, the reduction in correlations was slight. The average (r to z transformed) ability score stability correlation dropped from .842 to .838 in the Stradaptive 1 data and from .732 to .709 for Stradaptive 2 (Table 6). The reduction in conventional test stabilities shown in Table 7 was equivalent to that found in the stradaptive tests even though initial ability estimates were not able to inflate the zero-order conventional test stabilities. This suggests that the artifactual inflating effect of the intial ability estimates on stradaptive test stabilities was negligible.

Eta coefficients. No Stradaptive 1 score showed significant curvilinearity in the relationship between the test and retest score distributions. Of the three Stradaptive 2 scores with significant curvilinearity, no low order trends were apparent in the bivariate scatter plots. This suggests that the relationship between stradaptive initial test and retest scores is essentially linear.

Test-retest interval. Table 8 presents test-retest stability coefficients as a function of the length of the test-retest interval. With the exception of score 4, all Stradaptive 1 ability scores show monotonically decreasing stability with increasing time interval. The greatest decreases are observed for scores 2 and 5, the (N+1)*th* item and stratum scores. Scores 1, 3, 6 and 7 appear to be little affected by test-retest interval. Consistency scores 14 and 15 show increasing stability over time..

The 29-item analogous form of the conventional test had a considerably lower test-retest reliability (.828) than the best of the stradaptive scoring methods (score 8, r=.932) in the shortest (30-45 day) time interval. In the 61-79 day retest interval, the 29-item analogous form had a retest correlation of .860 while the retest stability of stradaptive score 8 was .848 and that of score 9 was .858. The 29-item analogous form corrected for internal consistency had a retest correlation of .916 in the 61-79 day interval, considerably higher than any of the stradaptive scores. Again however, the legitimacy of the numerous corrections involved in these data must be taken into account.

Although the distribution of test-retest intervals did not allow inclusion of Stradaptive 2 data in this analysis, an observation worthy of note is that the total group Stradaptive 2 stabilities were uniformly lower than the total group Stradaptive 1 stabilities (Table 7) even through the mean Stradaptive 2 test-retest interval was much shorter (24.6 vs. 47.9 days).

Correlations Between Stradaptive 1 and Conventional Test Scores

Table 9 shows product-moment correlations and eta coefficients of 40-item conventional test scores with Stradaptive 1 test scores. The highest correlations, those of scores 8 and 9 with the conventional test, when corrected for attenuation using test-retest stabilities were .942 and .938. Three of the Stradaptive 1 scores show significant curvilinearity in their relationship with the conventional test score. This is probably due to

Table 8

Test-retest Stability Correlations
for Stradaptive 1 and the Conventional Test
by Test-retest Interval

	Stradaptive 1			
		Test-Retest Interval		
	Total Group	31-45 Days	46-60 Days	61-79 Days
No. of Testees[a]	170	68	84	18
No. of Days				
Mean	47.870	40.333	51.311	64.722
SD	8.931	3.202	3.341	2.608
Skew	-.401	-.347	.449	.661
Ability Scores				
1	.837	.842	.818	.802
2	.786	.829	.781	.591
3	.844	.850	.841	.825
4	.833	.857	.783	.853
5	.787	.828	.792	.550
6	.841	.851	.832	.825
7	.872	.885	.859	.858
8	.919	.932	.914	.848
9	.911	.923	.908	.809
10	.841	.853	.835	.816
Consistency Scores				
11	.569	.503	.608	.587
12	.496	.555	.457	.513
13	.251	.285	.178	.354
14	.327	.250	.330	.553
15	.321	.236	.331	.540

	Conventional Test			
		Test-retest Inverval		
	Total Group	31-45 Days	46-60 Days	61-79 Days
No. of Testees[a]	194[b]	28	130	35
No. of Days				
Mean	53.567	41.750	53.469	64.743
SD	8.149	1.266	3.611	4.010
Skew	-.808	-.077	.073	1.855
40-item test	.913	.905	.924	.879
29-item analogous form[c]	.874	.828	.886	.860
29-item analogous form corrected for internal consistency	.931	.882	.943	.916

[a]Number of testees with valid data. Correlations for some scoring methods are based on fewer testees.

[b]Within group Ns do not add to total group N because of a single case in the 0-15 day inverval.

[c]All statistics for the 29-item analogous form are means of statistics from five combinations of items.

different scalings of the scoring methods.

Table 9

Correlations of Stradaptive 1 Scores with Conventional Test Scores (N=201)

Score	Correlation	Eta
Ability Scores		
1	.782	.800
2	.768	.799*
3	.790	.794
4	.784	.795
5	.769	.773
6	.791	.798
7	.812	.820
8	.859	.880*
9	.860	.885*
10	.800	.823
Consistency Scores		
11	.058	.303
12	.170	.355
13	.231	.346
14	.205	.531*
15	.200	.240

Note: The N reported refers to the maximum number of testees with valid data. Correlations for some scoring methods are based on fewer testees.
*Curvilinearity significant at $p<.05$

Table 9 also shows correlations of the stradaptive consistency scores with the conventional test scores. These correlations were uniformly low, ranging from .058 for score 11 to .231 for score 13. One consistency score showed a significant curvilinear relationship with the conventional test score. These data suggest that consistency scores provide information about testees which is not contained in ability scores derived from the conventional test.

Further Analyses of the Stradaptive Tests

Intercorrelations among Scores

Product-moment intercorrelations among the fifteen stradaptive scores are presented in the lower triangles of Table 10. In the Stradaptive 1 data, it is apparent that all the ability scores correlated highly among themselves (r=-.168 to .532). Closer inspection of the ability score intercorrelations revealed four relatively distinct clusters. The three item difficulty scores, scores 1, 2, and 3 formed three two-variable clusters, each

Table 10

Intercorrelations (Lower Triangle) and Inter-etas (Upper Triangle) Among Stradaptive Scores

Stradaptive 1 (N=476)

	Ability Scores										Consistency Scores				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1		914*	911*	996*	916*	912*	910*	906*	873*	911*	458*	564*	551*	539*	536*
2	844		970*	872*	996*	973*	972*	922*	913*	972*	334*	333	586*	604*	601*
3	880	948		869*	958	996	993	922*	946	998	300*	347	641*	637*	633*
4	986	831	847		858*	870*	873*	890*	846*	869*	409*	535	415	380	377
5	849	994	956	835		964*	961*	889*	922*	963*	205*	256	503*	510*	506*
6	881	951	996	854	959		994	911*	943*	998	213*	278	557*	547*	543*
7	882	950	993	854	957	995		953*	972*	996*	393*	415*	699*	744*	740*
8	895	877	910	875	882	908	933		985	922	453*	420*	442*	357	357
9	844	907	942	818	917	939	955	984		958*	448*	355	588*	540*	539*
10	881	949	998	851	957	999	995	913	944		316	385	640*	640*	637*
11	286	-033	047	320	040	059	028	-055	-168	051		832	555*	525	525
12	514	222	245	520	226	247	219	149	049	244	832		625*	597*	598*
13	430	461	534	399	464	532	494	228	240	530	401	505		975*	975*
14	394	458	512	372	466	522	457	172	173	517	424	596	960		999
15	392	457	511	370	466	521	455	170	172	516	423	497	960	999	

Stradaptive 2 (N=113)

	Ability Scores										Consistency Scores				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1		759	846	986	774	846	845	893	822	849	525	611	474	503	496
2	743		884	756	959	894	893	806	852	884	290	386	536	606	602
3	842	886		819	882	996	989	872	934	997	381	418	511	649*	644*
4	986	734	817		760*	826	826	871	926	830	493	591	434	356*	499*
5	747	988	895	734		902	902	802	848	902	190	284	514	527	520
6	841	894	996	820	902		993	873	928	998	316*	383*	525	579	574
7	841	893	989	821	902	993		896	945	991	394*	422	495	701*	695*
8	893	781	872	871	789	867	894		958	878	363	377	383	356	346
9	822	839	934	792	848	928	945	958		930	455*	431*	529*	438*	430*
10	843	883	997	821	902	998	991	873	930		392*	419	563	634	627
11	441	117	178	471	118	126	108	115	064	179		902*	610*	548*	549*
12	579	206	282	591	202	248	226	240	186	282	876		631*	599*	597*
13	387	472	479	406	474	484	442	155	168	484	400	458		944	942
14	429	520	550	443	522	573	501	199	177	556	387	455	944		999
15	423	516	544	437	517	568	495	191	169	550	385	455	942	999	

Note. Decimal points omitted. The N reported refers to the maximum number of subjects with valid data; correlations between some methods may be based on fewer subjects.

*Curvilinearity significant at $p<.05$

with their respective stratum difficulty scores, 4, 5, and 6. In addition, scores 3 and 6, the highest non-chance item and stratum scores, formed a tight cluster with scores 7 and 10, the interpolated stratum difficulty and average highest non-chance item scores. The common feature tying these latter four scores together seems to be their close reliance on the ceiling stratum. Only these four ability scores make explicit use of the ceiling stratum via the functionally related highest non-chance stratum. Score 9 also is dependent on the ceiling stratum but its further dependence on the basal stratum, and the fact that it is an average difficulty score, apparently lower its relationship with this cluster.

The final ability score cluster was composed of scores 8 and 9. The bond between these two scores appears to be that they are both average difficulty scores; score 8 is the overall average difficulty of all items correct and score 9 is the average difficulty of all items correct between the ceiling and basal strata.

Clustering among the consistency scores was even more obvious. Scores 11 and 12, the overall variability scores, formed one distinct cluster and scores 13, 14 and 15, reflecting the testee's variability between ceiling and basal strata, formed another.

Eta coefficients are shown in the upper triangles of Table 10. Although eta is an asymmetric statistic, to conserve computer time etas were calculated in one direction only, the rows being the independent variables. Due to the large sample size, curvilinearity was significant in most cases. But the actual differences between eta and r were small and the same clusters of scores emerged, thus yielding the same conclusions as the product-moment correlations regarding the similarities of scoring methods.

The intercorrelations and inter-etas for Stradaptive 2 were somewhat smaller than those of Stradaptive 1, but the same pattern observed in the Stradaptive 1 intercorrelations was apparent. Fewer significantly curvilinear relationships were observed for Stradaptive 2 but this is due to the smaller number of testees in the Stradaptive 2 analysis.

Utility of the Stradaptive Consistency Scores in Predicting Stability

Table 11 shows the test-retest correlations for scores on the stradaptive and conventional tests as a function of consistency score intervals computed from initial stradaptive test records. Retest correlations are shown separately for: 1) consistency score 11, each testee's standard deviation of difficulties of items encountered; 2) score 12, the standard deviation of items answered correctly; and 3) score 13, the standard deviation of difficulties for items answered correctly between the ceiling and basal strata.

Table 11 shows a strong moderator effect on test-retest reliability for consistency score 11, and, to a lesser extent for score 12, with no general moderator effect for score 13. For consistency score 11, the strongest moderator effect was observed for ability score 1. On this score, the very high consistency group (mean consistency score of .517) had a test-retest correlation of r=.940. As consistency decreased, test-retest reliability also decreased monotonically, with the very low consistency group (mean=1.038)

Table 11

Stradaptive 1 and Conventional Test Test-retest Correlations as a Function of Initial Test Consistency Scores 11, 12 and 13

	Status on Consistency Score 11				
	Very High	High	Average	Low	Very Low
Mean Consistency Score	.517	.625	.706	.815	1.038
Number of Testees in Interval[a]	27	30	41	43	29
Stradaptive Ability Score: 1	.940	.849	.847	.768	.652
2	.875	.721	.799	.778	.751
3	.956	.813	.878	.826	.708
4	.934	.840	.847	.731	.664
5	.896	.722	.793	.756	.741
6	.950	.798	.886	.820	.704
7	.970	.844	.902	.851	.758
8	.981	.927	.915	.853	.869
9	.983	.939	.907	.899	.889
10	.951	.792	.882	.822	.718
Conventional Test	.979	.890	.918	.826	.878

	Status on Consistency Score 12				
	Very High	High	Average	Low	Very Low
Mean Consistency Score	.379	.550	.656	.752	.955
Number of Testees in Interval[a]	30	39	40	27	34
Stradaptive Ability Score: 1	.892	.833	.909	.784	.724
2	.764	.778	.850	.823	.684
3	.913	.835	.900	.856	.697
4	.895	.813	.903	.715	.781
5	.783	.743	.870	.831	.670
6	.908	.827	.890	.867	.686
7	.943	.859	.921	.870	.737
8	.959	.920	.946	.841	.857
9	.968	.935	.926	.876	.883
10	.906	.823	.894	.858	.700
Conventional Test	.962	.852	.952	.620	.904

	Status on Consistency Score 13				
	Very High	High	Average	Low	Very Low
Mean Consistency Score	.119	.282	.376	.488	.670
Number of Testees in Interval[a]	34	17	29	30	31
Stradaptive Ability Score: 1	.853	.741	.804	.812	.825
2	.775	.755	.776	.876	.773
3	.746	.861	.871	.885	.810
4	.851	.800	.767	.801	.838
5	.758	.767	.780	.877	.812
6	.750	.873	.862	.893	.783
7	.790	.890	.906	.906	.822
8	.921	.930	.930	.924	.915
9	.892	.855	.947	.921	.912
10	.746	.873	.857	.887	.808
Conventional Test	.908	.765	.914	.926	.856

[a]Total number of testees with valid data. Retest reliabilities for some scoring methods are based on fewer cases.

obtaining a test-retest reliability of only r=.652. Similar results were obtained for the other ability scores using consistency score 11 as a moderator variable. The potential utility of score 11 is shown by the extremely high test-retest correlations for the very high consistency group for ability scores 8 and 9 (r=.981 and .983, respectively). This indicates that the retest ability scores of testees whose response records show little variability on initial testing are almost perfectly predictable from their initial ability test scores. Using these same ability scores and score 11 as the consistency score, the very low consistency group is considerably less predictable on retest (r=.869 and .889, respectively).

When testees were subgrouped on initial consistency from stradaptive score 11 and retest reliabilities were computed using conventional test scores, the results were not of the same pattern. Although the retest reliabilities were highest for the very high consistency subgroup (r=.979), the very low consistency group did not have the lowest reliability (r=.878). However, the fact that the very high consistency group had a very high test-retest reliability suggests some generality to the moderator effect for consistency scores as measured by score 11.

Consistency score 12 showed a meaningful moderator effect for a number of the ability scores. With the exception of ability scores 2, 4, 5 and 9 the very high consistency group had the highest test-retest reliability, and the very low consistency group had the lowest correlations. In no case, however, was there the monotonically decreasing reliability coefficients obtained for several scoring methods using consistency score 11 as the moderator variable; for score 12, the average consistency group tended to obtain a higher reliability than the high consistency group. For ability score 9, the only deviation from the monotonic trend was the low consistency group (r=.883). No general trend in stability correlations was observed for the scores on the conventional test when moderated by consistency score 12. Conventional score stabilities were, however, quite high (r=.962) for the very high consistency subgroup, as was found for score 11.

Score 13 functioned very poorly as a moderator of test-retest stability. For only two of the ability scores (1 and 4) was the test-retest correlation highest in the very high consistency group. For the majority of the other ability scores, and for the conventional test, stability correlations were highest for the low consistency subgroup.

Appendix Table A-4 shows test-retest reliability correlations as a function of consistency score intervals for Stradaptive 2. For this variation of the stradaptive test, the predicted pattern of stabilities did not occur for any of the consistency scores using any of the ability scores. These results could be due to sampling fluctuations, or they could be due to the differences between branching strategies used in Stradaptive 1 and 2.

Stability of the Stradaptive Test Response Records

Stability of within strata data. Table 12 presents the redundancy analyses for total number of items answered within strata and proportion correct within strata. The latter data were referred to as "subject

characteristic curves", possibly reflecting the scalability of the individual with respect to a set of items.

Table 12

Redundancy Analysis for Number of Items Administered and Proportions Correct Within Strata

Stradaptive 1	
Number of Items Administered within Strata	
Redundancy of retest given initial test	.414
Redundancy of initial test given retest	.439
Proportions Correct within Strata	
Redundancy of retest given initial test	.670
Redundancy of initial test given retest	.668
Stradaptive 2	
Number of Items Administered within Strata	
Redundancy of retest given initial test	.319
Redundancy of initial test given retest	.351
Proportions Correct within Strata	
Redundancy of retest given initial test	.528
Redundancy of initial test given retest	.471

For Stradaptive 1, 41.4% of the variance in number of items administered within strata was predictable on retest from initial test data. The proportion correct within strata was more predictable on retest, however. For these data, 67% of the retest variance was predictable from initial test scores. This result is equivalent to an average multiple correlation of about .82 in predicting an individual's proportion correct within a stratum at retest from his initial test data.

For Stradaptive 2, redundancy for number of items administered within strata was .319, while that for proportion correct was .528. These results support earlier findings that scores on Stradaptive 2 are less reliable than those on Stradaptive 1. However, it supports the finding with Stradaptive 1 that the proportions correct data are likely to be more useful than the number of items administered within strata.

Stability of total number administered. Test-retest correlations of total number of items administered in the two stradaptive tests were r=.335 and r=.055 for Stradaptive 1 and 2, respectively. This finding partially accounts for the fact that proportions correct were more stable than number of items administered within strata.

Relative Difficulties of Items Producing Different Kinds of Responses

Table 13 gives means and standard deviations of scores in both normal ogive difficulty units and stratum units for deviations of item difficulties from ability as determined from score 8. These distributions are presented as conditional on the type of response given (i.e., correct, incorrect or question mark).

The average deviation of items answered correctly from score 8 was 0.0, but this is artifactual since score 8 was defined as the mean difficulty of all items answered correctly. From the standard deviations in stratum units, it can be seen that the difficulties of 95% of all items answered correctly fell within 2.25 strata above or below the final ability estimate. It is also apparent that the difficulties of incorrect items were slightly greater than one stratum more difficult than correct responses.

Table 13

Deviations of Item Difficulties from Score 8 for Three Types of Response

	Stradaptive 1				Stradaptive 2			
	Normal Ogive Difficulty Units		Stratum Units		Normal Ogive Difficulty Units		Stratum Units	
Response	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Correct	.000	.745	.000	1.146	.000	.733	.000	1.128
Incorrect	.724	.751	1.114	1.155	.689	.730	1.060	1.123
Question Mark	.792	.775	1.218	1.192	.814	.791	1.252	1.217

Mean difficulties of items responded to with a question mark were greater than mean difficulties of items answered incorrectly--.068 and .125 units (or .11 and .19 strata) more difficult for the two stradaptive forms, respectively. Since the stradaptive testing strategy attempts to adapt the item difficulty to the ability of the testee, and since the question mark responses appear to indicate that items responded to in that way are even more difficult than those items answered incorrectly, it is obvious that the testee should be branched to a less difficult item following a question mark response. Discarding these item responses is an inefficient strategy for dealing with question mark responses.

Test Length vs. Ability

Table 14 shows correlations of all scores with test length (number of items administered to each testee at initial test) on both stradaptive forms. Also shown is the correlation of conventional test score with Stradaptive 1

test length. All stradaptive ability scores except scores 8 and 9, the average difficulty scores, correlated slightly with test length, on both forms of the stradaptive test. Consistency scores showed moderate to high correlations with test length on both forms. Stradaptive scores 8 and 9, and conventional test scores, had essentially zero correlations with test length in all cases.

Table 14

Correlation of Stradaptive and Conventional Test Scores with Number of Items Administered on the Stradaptive Test

Score	Stradaptive 1 (N=476)	Stradaptive 2 (N=113)
Ability Scores		
1	.263	.379
2	.254	.302
3	.292	.399
4	.258	.381
5	.253	.298
6	.301	.405
7	.241	.332
8	-.020	.094
9	-.046	.046
10	.293	.398
Consistency Scores		
11	.449	.375
12	.455	.415
13	.727	.693
14	.787	.781
15	.788	.782
Conventional Test	.037	....

Note. N's shown are number of testees with valid data. Certain scoring methods have fewer valid cases.

The most parsimonious explanation for the slight correlations of ability scores with test length is that the less discriminating items at the upper strata cause greater variation in the test record, which in turn causes the test to take longer to satisfy the termination criterion. This explanation is supported by the correlations of consistency scores with test length. This explanation fails, however, to explain the zero correlations of scores 8 and 9 with test length.

An alternative explanation is that test length is increased by inconsistent response records (due to test-testee interaction) which have a ceiling stratum more distant from actual ability level only because the range of item difficulties encountered was greater. Average difficulty scores would

not be affected by this phenomenon, but maximum performance scores (e.g., scores 1 and 4) and scores dependent upon the ceiling stratum (e.g., scores 3, 6, 7 and 10) would be. This explanation is supported by 1) the data showing positive correlations between test length and consistency scores (especially the indices of distance between the ceiling and basal strata, scores 14 and 15), and 2) the score intercorrelation data, which showed moderate positive correlations between consistency scores and all ability scores, except scores 8 and 9 which correlated only slightly with consistency scores. This explanation also suggests that there is an undesirable interaction between scoring method and the termination criterion, rather than any deficiences in the item pool

SUMMARY AND CONCLUSIONS

The most interesting distributional difference found in this study was that Stradaptive 2 scores had lower means than comparable Stradaptive 1 scores. This was surprising because Stradaptive 2 allowed testees to skip the more difficult items. Other distributional characteristics worthy of note were the close distributional similarities between scores on the 40-item conventional test and scores on the five 29-item analogous pairs of conventional tests, and the positive skew present in the stradaptive tests but not in the conventional test.

Although the distribution of underlying ability is not known for this college population, it is not unreasonable to assume that it is not normally distributed. Rather, a college population is likely to be positively skewed in verbal ability, since low ability testees would not qualify on entrance examinations which are highly verbal. This suggests that the stradaptive test better reflects the distribution of verbal ability in these testees, than does the conventional test. Furthermore, regardless of the distribution of ability in the population, the positive skew in the stradaptive scores suggests the capability of making finer discriminations among high ability testees than does the conventional test. And, it would be expected that a college population would include a number of very high ability testees whose scores would skew the distribution in a positive direction.

For inter-strategy comparisons it is important that internal consistencies of all tests be equal. In this study, the stradaptive tests had more discriminating items and thus higher internal consistencies.

Additional differences between the stradaptive and conventional tests confounded the comparison of test-retest stabilities between tests. Differences in lengths and memory effects were corrected for by creation of analogous conventional test pairs matching the stradaptive tests on psychometric characteristics. But the many corrections required limit the conclusions that can be drawn regarding stability comparisons. The results were rather inconclusive with the corrected conventional test stability being slightly higher than the best scores of Stradaptive 1.

Stradaptive 1 ability scores showed a decreasing trend in stability with time while stabilities for the conventional test showed no trend with time. However, the differences between time intervals were short and a longer

interval would be expected to show trends for both tests. The fact that the stabilities of the Stradaptive 1 scores did change with time while those of the conventional test did not may have been a function of the greater potential for memory effects to inflate test-retest stabilities in the conventional test, in which all 40 items were repeated on retest. The systematically decreasing trend of test-retest stabilities has not been observed in other empirical studies of adaptive tests (e.g., Betz and Weiss, 1974; Larkin and Weiss, 1974).

In contrast to the few meaningful inter-strategy comparisons provided by this study, much was learned about how to build a stradaptive test. Intercorrelations between the stradaptive scores showed four relatively distinct ability score clusters and two consistency score clusters. The ability score clusters included: 1) maximum performance scores; 2) (N+1)*th* item and stratum scores, 3) highest non-chance scores; and 4) average difficulty scores. Average difficulty scores showed the highest stabilities in this study. (N+1)*th* item and stratum scores showed the lowest stabilities, a finding probably due to the small number of strata available which allowed scores to change by 20% of the total score range on the basis of the response to the last item administered. The moderate stabilities of such chance-influenced scores as the highest difficulty scores is favorable in that it shows that the stradaptive test can contain an individual testee within items in the range of his ability. That the scores dependent on the ceiling stratum were only moderately stable may have been due to their joint dependence on central tendency and variability.

The consistency scores clustered into overall variability and distance between ceiling and basal strata scores. Score 11, an overall variability score, functioned as a meaningful moderator of ability score stabilities for Stradaptive 1. However, the stabilities of the consistency scores were only moderately high, although there was a tendency for the stabilities to increase with longer time intervals. Stradaptive 2 consistency scores were not predictive of test-retest stability of ability scores, but these results were given little weight because of the other erratic results obtained from Stradaptive 2.

Stabilities of the total stradaptive response records, as assessed by the redundancy analyses, were somewhat lower than the stabilities of the best ability scores but higher than those of the consistency scores. The proportions correct within strata were more stable than the numbers administered within strata. Total number of items administered in the stradaptive test was relatively unstable on retest.

The relative difficulties of correct, incorrect, and question mark responses suggest that the question mark response is used by the testee not when the testee is unsure of the correct response but rather when he has no idea what the correct response is and prefers not to guess randomly. Items which were responded to with a question mark were even more difficult, on the average, than those which the testee answered incorrectly. This result, when combined with the stability data for the two forms of stradaptive tests, suggests Stradaptive 1, in which question mark responses were counted as incorrect responses, is a better testing strategy than Stradaptive 2, in which question mark responses were ignored.

A slight correlation was obtained between test length and all ability scores except the average difficulty scores. This was explained as resulting from scores being defined as a joint function of central tendency and variability, the latter of which causes test length to increase. To confirm this hypothesis, a computer simulation of this phenomenon is required. If test scores can be changed by increasing the variability of the response record, which will probably be manipulated by changing the item discrimination, all but average difficulty scoring methods (i.e., score 8 and 9) will have to be discarded.

This study did not provide a clear answer to the question of interest: whether a conventional or a stradaptive testing strategy provides better measurement. A future study aimed at answering this question must control extraneous variables such as item discrimination, memory effects, effects of initial ability estimates and test length.

Two interesting aspects of the stradaptive testing strategy were not investigated in this study. First, the usefulness of an initial ability estimate was not determined. Monte carlo simulation methods would be appropriate to determine how much information is added to a stradaptive test by initial ability estimates when the ability estimates have differing degrees of correlation with underlying ability. The other important aspect of the stradaptive test that was not investigated in this study was the utility of flexible termination. What must be determined in future study is what variable, if any, is sufficiently related to error of measurement to provide a basis for deciding when to terminate the stradaptive test.

A final refinement of the stradaptive test that needs investigation is the development of an optimal scoring strategy for the stradaptive test. The average difficulty scores were the most stable in this study but a scoring technique based on a more adequate theoretical rationale might have superior psychometric characteristics, or more practical utility.

REFERENCES

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. (AD 768993).

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

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, 1974. (Ad 773961).

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

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

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, 1974. (AD 781894).

Meehl, P.E. Nuisance variables and the ex post facto design. In Feigl, H. & Maxwell, G. (Eds.), Minnesota studies in the philosophy of science. (Vol. 4). Minneapolis: University of Minnesota Press, 1970.

Mussio, J.J. A modification to Lord's model for tailored tests. Unpublished doctoral dissertation, University of Toronto, 1973.

Novick, M.R. Bayesian methods in psychological testing. Research Bulletin RB-69-31. Princeton, N.J.: Educational Testing Service, 1969.

Owen, R.J. A Bayesian approach to tailored testing. Research Bulletin RB-69-92. Princeton, N.J.: Educational Testing Service, 1969.

Stewart, D. & Love, W. A general canonical correlation index. Psychological Bulletin, 1968, 70, 160-163.

Waters, B.K. An empirical investigation of the stradaptive testing model for the measurement of human ability. Unpublished doctoral dissertation, Florida State University, 1974.

Waters, B.K. An empirical investigation of Weiss' stradaptive testing model. Paper presented at the Conference on Computerized Adaptive Testing, Washington, D.C., June 1975.

Weiss, D.J. Canonical correlation analysis in counseling psychology research. Journal of Counseling Psychology, 1972, 19, 241-252.

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

Weiss, D.J. Strategies of adaptive ability measurement. Research Report 74-5, Psychometric Methods Program, Department of Psychology, University of Minnesota, 1974. (AD A004270).

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

APPENDIX

Supplementary Tables

Table A-1

Normal Ogive Item Difficulty (b) and Discrimination (a) Item Parameters for Stradaptive 1 and 2

Stratum 1		Stratum 2		Stratum 3		Stratum 4		Stratum 5		Stratum 6		Stratum 7		Stratum 8		Stratum 9	
b	a	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b	a
-2.415	3.000	-1.989	1.756	-1.509	1.396	-.703	1.822	-.054	1.366	.728	.980	1.071	.718	1.893	.851	2.949	.840
-2.415	3.000	-1.779	1.536	-1.233	1.347	-.734	.917	.144	1.068	.337	.912	1.155	.638	2.033	.638	2.467	.475
-2.453	3.000	-2.216	1.524	-1.084	1.230	-.524	.862	-.132	.983	.651	.774	1.487	.618	1.928	.573	2.615	.434
-2.453	3.000	-1.679	1.460	-1.332	1.155	-.683	.861	.151	.967	.788	.703	1.334	.601	2.313	.540	2.857	.425
-2.716	3.000	-1.869	1.426	-1.636	1.081	-.592	.826	-.082	.908	.791	.627	1.535	.582	1.788	.505	2.351	.418
-2.716	3.000	-1.922	1.230	-1.342	1.020	-.746	.820	.161	.865	.486	.561	1.108	.564	2.045	.493	2.666	.416
-2.716	3.000	-1.880	1.137	-1.095	.986	-.567	.766	-.207	.865	.423	.550	1.395	.549	1.790	.486	2.320	.380
-2.665	1.790	-2.127	1.097	-1.648	.922	-.851	.748	-.254	.862	.977	.524	1.171	.518	1.880	.446	2.368	.345
		-1.636	*1.081*														
-2.535	1.587	-2.225	1.071	-1.421	.920	-.473	.710	.208	.858	.368	.505	1.298	.515	2.070	.434	3.113	.325
-2.807	1.482	-1.672	1.023	-1.207	.905	-.399	.681	.165	.826	.968	.490	1.376	.512	2.132	.421	2.504	.214
-2.460	1.289	-1.710	.986	-1.057	.890	-.905	.671	-.228	.810	.461	.486	1.440	.487	2.307	.400		
-2.776	1.255	-2.262	.977	-1.341	.886	-.998	.671	.300	.776	.457	.483	1.307	.440	1.649	.391		
-2.469	1.155	-2.208	.961	-1.308	.871	-.690	.665	.172	.774	.650	.480	1.246	.427	1.819	.362		
-2.434	1.008	-1.657	.927	-1.653	.822	-.813	.665	.172	.772	.784	.448	1.004	.418	2.265	.352		
		-1.648	*.922*														
		-1.653	*.822*														
-2.865	1.008	-2.322	.796	-1.100	.770	-.562	.662	.075	.756	.708	.443	1.005	.405	2.179	.339		
-2.943	.956	-1.795	.774	-1.554	.768	-.581	.662	-.285	.750	.652	.431	1.263	.387				
-2.833	.943	-1.804	.623	-1.068	.760	-.839	.653	.136	.697	.615	.408	1.151	.383				
-2.737	.933	-1.934	.740	-1.433	.760	-.739	.647	.240	.664	.976	.377	1.359	.371				
-2.884	.912	-2.285	.696	-1.405	.748	-.630	.647	.173	.637	.829	.372	1.240	.360				
-2.538	.879	-1.827	.660	-1.147	.727	-.850	.642	-.184	.627	.747	.372	1.598	.349				
						.542	*.259*										
-2.554	.788	-1.745	.627	-1.418	.714	-.480	.638	-.281	.620	.920	.369	1.210	.346				
-2.810	.742	-1.699	.590	-1.627	.710	-.404	.637	.246	.609	.977	.310	1.473	.341				
-2.499	.685	-2.191	.558	-1.472	.667	-.730	.627	.001	.607			1.613	.341				
-2.817	.672	-1.892	.515	-1.603	.659	-.719	.609	-.281	.606			1.630	.322				
-2.540	.669	-2.196	.505	-1.331	.623	-.525	.602	-.296	.579			1.357	.312				
-2.498	.662	-1.711	.468	-1.037	.577	-.935	.596	-.248	.571								
-2.509	.637	-2.211	.439	-1.174	.571	-.413	.588	-.215	.562								
-2.393	.615	-2.082	.422	-1.269	.562	-.680	.682	.329	.527								
-2.578	.570	-1.804	.418	-1.074	.555	-.725	.568	-.233	.505								
-2.980	.559	-1.825	.417	-1.021	.538	-.835	.562	-.319	.501								
-2.732	.519	-2.120	.407	-1.013	.524	-.784	.543	-.078	.501								
-2.769	.497	-1.921	.320	-1.307	.524	-.889	.533	-.035	.474								
-2.675	.476	-1.840	.313	-1.300	.521	-.686	.512	-.171	.468								
-2.559	.443			-1.187	.519	-.956	.482	.188	.436								
-2.946	.406			-1.568	.487	-.525	.480	-.233	.434								
				-1.265	.440	-.576	.476	.089	.428								
				-1.594	.383	-.617	.472	.149	.419								
				-1.348	.398	-.395	.405	.189	.417								
				-1.080	.317	-.363	.402	-.086	.410								
						-.738	.400	-.257	.400								
						-.581	.397	.076	.387								
						-.376	.379	.086	.371								
						-.896	.338	-.045	.351								
						-.927	.332	-.125	.317								
						-.343	.323										
						-.673	.301										

Note: Italicized items appeared only in the Stradaptive 1 item pool. Underlined items appeared only in the Stradaptive 2 item pool. All others appeared in both item pools.

Table A-2

Conventional Test Item Parameters

	Traditional		Normal Ogive	
	p	r_{bis}	b	a
Mean	.537	.472	-.188	.543
S.D.	.101	.078	.592	.112
Maximum	.661	.612	1.155	.774
Minimum	.267	.296	-.956	.310
	.661	.434	-.956	.482
	.656	.543	-.739	.647
	.659	.490	-.835	.562
	.469	.572	.136	.697
	.646	.520	-.719	.609
	.646	.477	-.784	.543
	.651	.531	-.730	.627
	.640	.494	-.725	.568
	.634	.543	-.630	.647
	.634	.503	-.680	.582
	.623	.456	-.686	.512
	.558	.518	-.281	.606
	.608	.371	-.738	.400
	.613	.320	-.896	.338
	.607	.516	-.525	.602
	.615	.315	-.927	.332
	.604	.427	-.617	.472
	.602	.538	-.480	.638
	.458	.612	.172	.774
	.458	.611	-.172	.772
	.557	.448	-.319	.501
	.559	.501	-.296	.579
	.559	.527	-.281	.620
	.549	.496	-.248	.571
	.542	.451	-.233	.505
	.539	.531	-.184	.627
	.542	.490	-.215	.562
	.529	.424	.171	.468
	.471	.385	-.189	.417
	.514	.448	.078	.501
	.500	.519	-.001	.607
	.506	.428	.035	.474
	.449	.520	.246	.609
	.470	.400	.188	.436
	.463	.537	.173	.637
	.340	.359	1.151	.383
	.267	.538	1.155	.638
	.386	.296	.977	.310
	.335	.440	.968	.490
	.365	.353	.976	.377

Table A-3

Characteristics of Score Distributions for Stradaptive 1, Stradaptive 2, and the Conventional Test on Retest

Stradaptive 1					
Score	N	Mean	S.D.	Skew	Kurtosis
Ability Scores					
1	180	1.250	1.172	-.334	-.635
2	180	.854	1.410	.107	-.760*
3	180	.793	1.286	-.066	-.822*
4	180	1.200	1.141	-.578*	-.539
5	180	.870	1.396	.010	-.671
6	180	.629	1.246	-.176	-.817*
7	180	.697	1.206	-.134	-.871*
8	180	.129	1.070	.017	-.785*
9	160	.276	1.124	.068	-.833*
10	180	.607	1.241	-.154	-.846*
Consistency Scores					
11	180	.746	.195	.630*	.232
12	180	.668	.228	.550*	.476
13	160	.398	.232	.421*	-.431
14	180	2.011	.932	.675*	.270
15	180	2.072	.921[a]	.662*	.233

Stradaptive 2					
Score	N	Mean	S.D.	Skew	Kurtosis
Ability Scores					
1	98	.809	1.126	.350	- .789
2	98	.324	1.283	.418	- .162
3	98	.302	1.215	.436	- .504
4	98	.767	1.094	.274	-1.015*
5	98	.315	1.261	.441	- .111
6	98	.171	1.205	.383	- .632
7	98	.256	1.153	.479	- .539
8	98	-.289	.956	.604*	- .282
9	83	-.137	.996	.667*	.235
10	97	.114	1.187	.359	- .586
Consistency Scores					
11	98	.761	.179	.407	.339
12	98	.677	.202	.291	- .269
13	83	.383	.212	-.031	-1.308*
14	98	1.898	.860	.217	- .933
15	98	1.918	.849[a]	.207	- .914

Conventional Test					
Score	N	Mean	S.D.	Skew	Kurtosis
40-items	194	.619	.211	-.195	- .796*
29-item analogous form[b]	194	.620	.213	-.166	- .870*

[a] S.D. is multiplied by .65.

[b] All statistics for the 29-item analogous form are means of statistics calculated on five combinations of items.

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

Table A-4

Stradaptive 2 Test-retest Correlations as a Function of Initial Test Consistency Scores 11, 12, and 13

		Status on Consistency Score 11				
		Very High	High	Average	Low	Very Low
Mean Consistency Score		.524	.631	.704	.820	1.033
Number of Testees in Interval[a]		18	14	15	15	17
Stradaptive Ability Score:	1	.376	.823	.747	.887	.598
	2	.007	.877	.535	.732	.588
	3	.604	.826	.680	.847	.603
	4	.219	.848	.733	.842	.592
	5	.151	.895	.558	.769	.586
	6	.550	.823	.669	.858	.534
	7	.592	.858	.698	.881	.563
	8	.589	.876	.873	.912	.764
	9	.846	.860	.786	.918	.662
	10	.569	.829	.689	.856	.621

		Status on Consistency Score 12				
		Very High	High	Average	Low	Very Low
Mean Consistency Score		.346	.552	.661	.765	.973
Number of Testees in Interval[a]		17	17	15	15	15
Stradaptive Ability Score:	1	.330	.769	.905	.867	.750
	2	.581	.590	.827	.440	.665
	3	.469	.803	.910	.766	.665
	4	.306	.723	.876	.844	.769
	5	.575	.632	.874	.647	.656
	6	.449	.798	.915	.688	.666
	7	.509	.819	.937	.700	.669
	8	.500	.874	.944	.846	.805
	9	.718	.809	.955	.785	.766
	10	.471	.808	.916	.783	.679

		Status on Consistency Score 13				
		Very High	High	Average	Low	Very Low
Mean Consistency Score		.117	.255	.379	.472	.665
Number of Testees in Interval[a]		13	14	13	13	14
Stradaptive Ability Score:	1	.590	.758	.771	.850	.848
	2	.652	.628	.605	.731	.717
	3	.773	.808	.835	.817	.795
	4	.555	.748	.783	.823	.876
	5	.562	.669	.642	.802	.714
	6	.774	.736	.787	.824	.801
	7	.685	.747	.792	.870	.797
	8	.732	.859	.807	.900	.903
	9	.715	.868	.805	.958	.822
	10	.771	.816	.827	.838	.799

[a]Total number of testees with valid data. Retest reliabilities for some scoring methods are based on fewer cases.

DISTRIBUTION LIST

Navy

4 Dr. Marshall J. Farr, Director
Personnel and Training Research Programs
Office of Naval Research (Code 458)
Arlington, VA 22217

1 ONR Branch Office
495 Summer Street
Boston, MA 02210
ATTN: Research Psychologist

1 ONR Branch Office
1030 East Green Street
Pasadena, CA 91101
ATTN: E.E. Gloye

1 ONR Branch Office
536 South Clark Street
Chicago, IL 60605
ATTN: M.A. Bertin

6 Director
Naval Research Laboratory
Code 2627
Washington, DC 20390

12 Defense Documentation Center
Cameron Station, Building 5
5010 Duke Street
Alexandria, VA 22314

1 Special Assistant for Manpower
OASN (M&RA)
Pentagon, Room 4E794
Washington, DC 20350

1 LCDR Charles J. Theisen, Jr., MSC, USN
4024
Naval Air Development Center
Warminster, PA 18974

1 Chief of Naval Reserve
Code 3055
New Orleans, LA 70146

1 Navy Personnel Research and
Development Center
Code 9041
San Diego, California 92152
Attn: Dr. J. D. Fletcher

1 Dr. Lee Miller
Naval Air Systems Command
AIR-413E
Washington, DC 20361

1 Commanding Officer
U.S. Naval Amphibious School
Coronado, CA 92155

1 Chief
Bureau of Medicine & Surgery
Research Division (Code 713)
Washington, DC 20372

1 Chairman
Behavioral Science Department
Naval Command & Management Division
U.S. Naval Academy
Luce Hall
Annapolis, MD 21402

1 Chief of Naval Education & Training
Naval Air Station
Pensacola, FL 32508
ATTN: CAPT Bruce Stone, USN

1 Mr. Arnold Rubinstein
Naval Material Command (NAVMAT 03424)
Room 820, Crystal Plaza #6
Washington, DC 20360

1 Commanding Officer
Naval Medical Neuropsychiatric
Research Unit
San Diego, CA 92152

1 Director, Navy Occupational Task
Analysis Program (NOTAP)
Navy Personnel Program Support
Activity
Building 1304, Bolling AFB
Washington, DC 20336

1 Dr. Richard J. Niehaus
Office of Civilian Manpower Management
Code 00A
Washington, DC 20390

1 Department of the Navy
Office of Civilian Manpower Management
Code 263
Washington, DC 20390

1 Chief of Naval Operations (OP-987E)
Department of the Navy
Washington, DC 20350

1 Superintendent
Naval Postgraduate School
Monterey, CA 93940
ATTN: Library (Code 2124)

1 Commander, Navy Recruiting Command
4015 Wilson Boulevard
Arlington, VA 22203
ATTN: Code 015

1 Mr. George N. Graine
Naval Ship Systems Command
SHIPS 047C12
Washington, DC 20362

1 Chief of Naval Technical Training
Naval Air Station Memphis (75)
Millington, TN 38054
ATTN: Dr. Norman J. Kerr

1 Dr. William L. Maloy
Principal Civilian Advisor
for Education & Training
Naval Training Command, Code 01A
Pensacola, FL 32508

1 Dr. Alfred F. Smode, Staff Consultant
Training Analysis & Evaluation Group
Naval Training Equipment Center
Code N-OOT
Orlando, FL 32813

1 Dr. Hanns H. Wolff
Technical Director (Code N-2)
Naval Training Equipment Center
Orlando, FL 32813

1 Chief of Naval Training Support
Code N-21
Building 45
Naval Air Station
Pensacola, FL 32508

1 Dr. Martin Wiskoff
Navy Personnel R&D Center
San Diego, CA 92152

5 Navy Personnel R&D Center
San Diego, CA 92152
ATTN: Code 10

1 D. M. Gragg, CAPT, MC, USN
Head, Educational Programs Development Department
Naval Health Sciences Education and Training Command
Bethesda, MD 20014

Army

1 Headquarters
U.S. Army Administration Center
Personnel Administration Combat Development Activity
ATCP-HRO
Ft. Benjamin Harrison, IN 46249

1 Armed Forces Staff College
Norfolk, VA 23511
ATTN: Library

1 Commandant
United States Army Infantry School
ATTN: ATSH-DET
Fort Benning, GA 31905

1 Deputy Commander
U.S. Army Institute of Administration
Fort Benjamin Harrison, IN 46216
ATTN: EA

1 Dr. Stanley L. Cohen
U.S. Army Research Institute
1300 Wilson Boulevard
Arlington, VA 22209

1 Dr. Ralph Dusek
U.S. Army Research Institute
1300 Wilson Boulevard
Arlington, VA 22209

1 Mr. Edmund F. Fuchs
U.S. Army Research Institute
1300 Wilson Boulevard
Arlington, VA

1 Dr. J.E. Uhlaner, Technical Director
U.S. Army Research Institute
1300 Wilson Boulevard
Arlington, VA 22209

1 HQ USAREUR & 7th Army
ODCSOPS
USAREUR Director of GED
APO New York 09403

Air Force

1 Research Branch
AF/DPMYAR
Randolph AFB, TX 78148

1 AFHRL/DOJN
Stop #63
Lackland AFB, TX 78236

1 Dr. Robert A. Bottenberg (AFHRL/CA)
Stop #63
Lackland AFB, TX 78236

1 Dr. Martin Rockway (AFHRL/TT)
Lowry AFB
Colorado 80230

1 Major P.J. DeLeo
Instructional Technology Branch
AF Human Resources Laboratory
Lowry AFB, CO 80230

1 AFOSR/NL
1400 Wilson Boulevard
Arlington, VA 22209

1 Commandant
USAF School of Aerospace Medicine
Aeromedical Library (SUL-4)
Brooks AFB, TX 78235

1 CAPT Jack Thorpe, USAF
Flying Training Division (HRL)
Williams AFB, AZ 85224

1 AFHRL/PE
Stop 63
Lackland AFB, TX 78236

Marine Corps

1 Mr. E.A. Dover
Manpower Measurement Unit (Code MPI)
Arlington Annex, Room 2413
Arlington, VA 20380

1 Commandant of the Marine Corps
Headquarters, U.S. Marine Corps
Code MPI-20
Washington, DC 20380

1 Director, Office of Manpower Utilizatio
Headquarters, Marine Corps (Code MPU)
MCB (Building 2009)
Quantico, VA 22134

1 Dr. A.L. Slafkosky
Scientific Advisor (Code RD-1)
Headquarters, U.S. Marine Corps
Washington, DC 20380

1 Chief, Academic Department
Education Center
Marine Corps Development and Education Command
Marine Corps Base
Quantico, VA 22134

Coast Guard

1 Mr. Joseph J. Cowan, Chief
Psychological Research Branch (G-P-1/62)
U.S. Coast Guard Headquarters
Washington, DC 20590

Other DOD

1 Lt. Col. Henry L. Taylor, USAF
Military Assistant for Human Resources
OAD (E&LS) ODDR&E
Pentagon, Room 3D129
Washington, DC 20301

1 Col. Austin W. Kibler
Advanced Research Projects Agency
Human Resources Research Office
1400 Wilson Boulevard
Arlington, VA 22209

1 Dr. Harold F. O'Neil, Jr.
Advanced Research Projects Agency
Human Resources Research Office
1400 Wilson Boulevard, Room 625
Arlington, VA 22209

1 Helga L. Yoich
Advanced Research Projects Agency
Manpower Management Office
1400 Wilson Boulevard
Arlington, VA 22209

Other Government

1 Dr. Lorraine D. Eyde
Personnel Research and Development Center
U.S. Civil Service Commission
1900 E. Street, N.W.
Washington, DC 20415

1 Dr. William Gorham, Director
Personnel Research and Development Center
U.S. Civil Service Commission
1900 E. Street, N.W.
Washington, DC 20415

1 Dr. Vern Urry
Personnel Research and Development Center
U.S. Civil Service Commission
1900 E. Street, N.W.
Washington, DC 20415

1 U.S. Civil Service Commission
Federal Office Bldg.
Chicago Regional Staff Div.
Attn: C. S. Winiewicz
Regional Psychologist
230 So. Dearborn St.
Chicago, IL 60604

Miscellaneous

1 Dr. Scarvia B. Anderson
Educational Testing Service
17 Executive Park Drive, N.E.
Atlanta, GA 30329

1 Dr. John Annett
The Open University
Milton Keynes
Buckinghamshire
ENGLAND

1 Dr. Richard Snow
Stanford University
School of Education
Stanford, CA 94305

1 Dr. Gerald V. Barrett
University of Akron
Department of Psychology
Akron, OH 44325

1 Dr. Bernard M. Bass
University of Rochester
Management Research Center
Rochester, NY 14627

1 Mr. Kenneth M. Bromberg
Manager - Washington Operations
Information Concepts, Inc.
1701 North Fort Myer Drive
Arlington, VA 22209

1 Dr. Norman Cliff
University of Southern California
Department of Psychology
University Park
Los Angeles, CA 90007

1 Century Research Corporation
4113 Lee Highway
Arlington, VA 22207

1 Dr. Kenneth E. Clark
University of Rochester
College of Arts & Sciences
River Campus Station
Rochester, NY 14627

1 Dr. Rene' V. Dawis
University of Minnesota
Department of Psychology
Minneapolis, MN 55455

1 Dr. Norman R. Dixon
Room 170
190 Lothrop Street
Pittsburgh, PA 15260

1 Dr. Robert Dubin
University of California
Graduate School of Administration
Irvine, CA 92664

1 Dr. Marvin D. Dunnette
University of Minnesota
Department of Psychology
Minneapolis, MN 55455

1 ERIC
Processing and Reference Facility
4833 Rugby Avnue
Bethesda, MD 20014

1 Dr. Victor Fields
Montgomery College
Department of Psychology
Rockville, MD 20850

1 Dr. Albert Glickman
American Institutes for Research
Foxhall Square
3301 New Mexico Avenue, N.W.
Washington, DC 20016

1 Dr. Ruth Day
Yale University
Department of Psychology
New Haven, CT 06520

1 Dr. Robert Glaser, Director
University of Pittsburgh
Learning Research & Development Center
Pittsburgh, PA 15213

1 Dr. Robert Vineberg
HumRRO Western Division
27857 Berwick Drive
Carmel, CA 93921

1 LCol CRJ Lafleur, Director
Personnel Applied Research
National Defence HQ
Ottawa, Canada K1A OK 2

1 Mr. Harry H. Harman
Educational Testing Service
Princeton, NJ 08540

1 Dr. Richard S. Hatch
Decision Systems Associates, Inc.
11428 Rockville Pike
Rockville, MD 20852

1 Dr. M.D. Havron
Human Sciences Research, Inc.
7710 Old Spring House Road
West Gate Industrial Park
McLean, VA 22101

1 HumRRO
Division No. 3
P.O. Box 5787
Presidio of Monterey, CA 93940

1 HumRRO
Division No. 4, Infantry
P.O. Box 2086
Fort Benning, GA 31905

1 HumRRO
Division No. 5, Air Defense
P.O. Box 6057
Fort Bliss, TX

1 HumRRO
Division No. 6, Library
P.O. Box 428
Fort Rucker, IL 36360

1 Dr. Lawrence B. Johnson
Lawrence Johnson & Associates, Inc.
200 S. Street, N.W., Suite 502
Washington, DC 20009

1 Dr. Steven W. Keele
University of Oregon
Department of Psychology
Eugene, OR 97403

1 Dr. David Klahr
Carnegie-Mellon University
Department of Psychology
Pittsburgh, PA 15213

1 Dr. Frederick M. Lord
Educational Testing Service
Princeton, NJ 08540

1 Dr. Ernest J. McCormick
Purdue University
Department of Psychological Sciences
Lafayette, IN 47907

1 Dr. Robert R. Mackie
Human Factors Research, Inc.
6780 Cortona Drive
Santa Barbara Research Park
Goleta, CA 93017

1 Mr. Edmond Marks
405 Old Main
Pennsylvania State University
University Park, PA 16802

1 Dr. Leo Munday, Vice-President
American College Testing Program
P.O. Box 168
Iowa City, IA 52240

1 Mr. A. J. Pesch, President
Eclectech Associates, Inc.
P.O. Box 178
North Stonington, CT 06359

1 Mr. Luigi Petrullo
2431 North Edgewood Street
Arlington, VA 22207

1 Dr. Diane M. Ramsey-Klee
R-K Research & System Design
3947 Ridgemont Drive
Malibu, CA 90265

1 Dr. Joseph W. Rigney
University of Southern California
Behavioral Technology Laboratories
3717 South Grand
Los Angeles, CA 90007

1 Dr. Leonard L. Rosenbaum, Chairman
Montgomery College
Department of Psychology
Rockville, MD 20850

1 Dr. George E. Rowland
Rowland and Company, Inc.
P.O. Box 61
Haddonfield, NJ 08033

1 Dr. Arthur I. Siegel
Applied Psychological Services
404 East Lancaster Avenue
Wayne, PA 19087

1 Dr. G. Harold Stone
1428 Virginia Avenue
Glendale, CA 91202

1 Mr. Dennis J. Sullivan
725 Benson Way
Thousand Oaks, CA 91360

1 Dr. Benton J. Underwood
Northwestern University
Department of Psychology
Evanston, IL 60201

1 Dr. Anita West
Denver Research Institute
University of Denver
Denver, CO 80210

Previous Reports in this Series

73-1. Weiss, D.J. & Betz, N.E. Ability Measurement: Conventional or Adaptive? February 1973 (AD 757788).

73-2. Bejar, I.I. & Weiss, D.J. Comparison of Four Empirical Differential Item Scoring Procedures. August 1973.

73-3. Weiss, D.J. The Stratified Adaptive Computerized Ability Test. September 1973 (AD 768376).

73-4. Betz, N.E. & Weiss, D.J. An Empirical Study of Computer-Administered Two-stage Ability Testing. October 1973 (AD 768993).

74-1. DeWitt, L.J. & Weiss, D.J. A Computer Software System for Adaptive Ability Measurement. January 1974 (AD 773691).

74-2. McBride, J.R. & Weiss, D.J. A Word Knowledge Item Pool for Adaptive Ability Measurement. June 1974 (AD 781894).

74-3. Larkin, K.C. & Weiss, D.J. An Empirical Investigation of Computer-Administered Pyramidal Ability Testing. July 1974 (AD 783553).

74-4. Betz, N.E. & Weiss, D.J. Simulation Studies of Two-stage Ability Testing. October 1974 (AD A001230).

74-5. Weiss, D.J. Strategies of Adaptive Ability Measurement. December 1974. (AD A004270).

75-1. Larkin, K.C. & Weiss, D.J. An Empirical Comparison of Two-stage and Pyramidal Adaptive Ability Testing. February 1975. (AD A006733).

75-2. McBride, J.R. & Weiss, D.J. TETREST: A FORTRAN IV program for calculating tetrachoric correlations. March 1975. (AD A007572).

75-3. Betz, N.E. & Weiss, D.J. Empirical and Simulation Studies of Flexilevel Ability Testing. July 1975.

AD Numbers are those assigned by the Defense Documentation Center, for retrieval through the National Technical Information Service.

Copies of these reports are available, while supplies last, from:

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