A SIMULATION STUDY OF STRADAPTIVE ABILITY TESTING

C. DAVID VALE

AND

DAVID J. WEISS

RESEARCH REPORT 75-6

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

DECEMBER 1975

Prepared under contract No. N00014-76-C-0243, NR150-382
with the
Personnel and Training Research Programs
Psychological Sciences Division
Office of Naval Research

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

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Research Report 75-6	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A Simulation Study of 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) N00014-76-C-0243
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.:NR150-382
11. CONTROLLING OFFICE NAME AND ADDRESS Personnel and Training Research Programs Office of Naval Research Arlington, Virginia 22217		12. REPORT DATE December 1975 13. NUMBER OF PAGES 51
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

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

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

18. SUPPLEMENTARY NOTES

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

testing	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 test and two forms of a stradaptive test were administered to thousands of simulated subjects by minicomputer. Characteristics of the three tests using several scoring techniques were investigated while varying the discriminating power of the items, the lengths of the tests, and the availability of prior information about the testee's ability level. The tests were evaluated in terms of their correlations with underlying ability, the amount of information they provided about ability, and the equiprecision of measurement they exhibited. Major findings were 1) scores on the conven-

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)

tional test correlated progressively less with ability as item discriminating power was increased beyond a=1.0; 2) the conventional test provided increasingly poorer equiprecision of measurement as items became more discriminating; 3) these undesirable characteristics were not characteristic of scores on the stradaptive test; 4) the stradaptive test provided higher score-ability correlations than the conventional test when item discriminations were high; 5) the stradaptive test provided more information and better equiprecision of measurement than the conventional test when test lengths and item discriminations were the same for the two strategies; 6) the use of valid prior ability estimates by stradaptive strategies resulted in scores which had better measurement characteristics than scores derived from a fixed entry point; 7) a Bayesian scoring technique implemented within the stradaptive testing strategy provided scores with good measurement characteristics; and 8) further research is necessary to develop improved flexible termination criteria for the stradaptive test.

CONTENTS

A Simulation Study of Stradaptive Ability Testing

Adaptive Testing

In constructing tests to measure mental abilities, the goal is to extract as much information as possible about an individual's ability level from a limited set of test items. Until recently, constrained to use paper and pencil administration techniques, test makers typically built a "peaked" ability test with all items of a difficulty such that a person with average ability could answer about half of them correctly. This type of test provided the highest level of information in the middle range of ability (Lord, 1970). Assuming a normal distribution of ability, the middle range is where most individuals were concentrated. Consequently, a perfectly peaked ability test provided the highest possible average level of information from a fixed set of items.

Unfortunately, for the individual with a high level of ability, a test peaked at the average ability level provided little information about his or her ability because all the items were too easy. The resulting test score for these individuals simply indicated that they were of high ability, but the test was not sufficiently sensitive to differentiate within the high ability group. Similarly, for individuals of low ability the test was often unable to provide meaningful measurement.

As on-line computer systems became widely available, it became possible to adapt a test to an individual, giving him/her items which were neither too difficult nor too easy regardless of his/her ability level. The goal of adaptive, or tailored, testing is to administer to the testee the subset of available items that will provide the maximum amount of information about his ability level. If his ability were known *a priori*, he would be given a set of items that a person of his ability level would answer about half correctly and about half incorrectly. Since a testee's ability is not accurately known before testing begins, the testing procedure must be designed to adapt to the individual's ability level as it is estimated during the course of testing, and thereby select the optimal item set for that individual.

A variety of approximation techniques, or strategies of item selection, have been suggested, ranging from simple two-stage techniques to complex Bayesian and maximum likelihood techniques (see Weiss, 1974, for a description and comparison of adaptive testing strategies). This paper reports on the stratified-adaptive or stradaptive (Weiss, 1973) testing strategy.

The Stradaptive Testing Strategy

The stradaptive test, as proposed by Weiss (1973), is a computer-based analogue of Binet's intelligence testing approach. It is based on an item pool composed of a collection of peaked tests, or strata, ordered by difficulty and equally spaced along the ability continuum. The best or most discriminating items are placed at the beginning of each stratum.

Using this strategy, an initial stratum assignment is made on the basis of some prior information about the testee's ability. Testing begins with the first item of this stratum. On the basis of the testee's response to each item, he is branched either up or down (typically by one stratum) to a more or less

difficult item. As in Binet's test, termination occurs when a ceiling stratum is reached; thus, the number of items administered to different individuals can vary. The ceiling stratum is one in which the testee answers all items incorrectly (or only a chance proportion correct if multiple-choice items are used). Weiss (1973) suggested, as an operational definition of this ceiling level, the least difficult stratum in which a chance proportion (or less) of items had been answered correctly after five items from that stratum had been administered. As with the Binet test, the stradaptive test also has a basal level which is defined as the most difficult stratum in which all items administered are answered correctly.

Scoring of the stradaptive test can result in ten scores which can be used to estimate ability level and five scores which reflect response consistency. The consistency scores are designed to reflect aspects of the interaction of an individual and a given item pool which might reflect the degree of error in the ability level scores. Examples of stradaptive response records, and further discussion of the logic of this testing strategy can be found in Weiss (1973), Vale and Weiss (1975) and Vale (1975).

Empirical studies of the stradaptive strategy. To date there have been two "live testing" or empirical studies of the stradaptive strategy. Vale and Weiss (1975) administered two forms of the stradaptive test and one conventional test to college students. All tests were administered by cathode ray terminals (CRTs) connected to a time-shared computer. The primary goals of this study were 1) comparison of the stradaptive test and a conventional test with respect to test-retest stability; 2) comparisons of the ten ability scores with respect to test-retest stability; and 3) investigation of the utility of the stradaptive consistency indices in predicting test-retest stabilities.

In the first part of the study, single administration data on the stradaptive test as originally proposed by Weiss (1973) were collected from 476 subjects and retest data were collected from 170 subjects. Analyses showed that meaningful comparisons between the test-retest reliabilities of the stradaptive and conventional strategies were precluded by many inequalities between the two tests. These included different test lengths, different proportions of items presented both on test and retest, unequal item discriminations, and the unknown influence of initial ability estimates on stability. After attempts to statistically correct for these inequities, no meaningful difference was found between strategies with respect to test-retest stability.

Intratest comparisons of stradaptive scores were informative, however. The ten ability scores grouped themselves into four clusters: 1) maximum performance scores, such as the difficulty of the most difficult item correct; 2) scores reflecting the difficulties of the next item or stratum the testee would have been given had the test continued for one more stage; 3) scores derived from the difficulty of the most difficult stratum in which the proportion correct was greater than chance responding; and 4) average difficulty scores. The average difficulty scores had the highest test-retest stabilities, and the scores derived from difficulties of hypothetical next items and strata had the lowest.

Three of the five consistency scores were evaluated in terms of their utility as variables moderating test-retest stability. This was done by subgrouping

testees on the basis of initial test consistency scores and comparing the test-retest stabilities of the groups. Two of the consistency scores--the standard deviations of difficulties of all items administered and of all items answered correctly--were quite predictive of test-retest stability, but the other consistency score was not.

In the second part of the study, a modification of the original stradaptive testing strategy was evaluated. In computerized stradaptive testing, testees may respond with a question mark if they do not know the correct answer to an item and prefer not to guess. In the original version of the stradaptive test, these "omits" were counted as incorrect, both in terms of branching decisions and in scoring. To evaluate the impact of not penalizing testees for honestly admitting they did not know the answer to an item, a modified version of the stradaptive test was studied. In this version, omissions were ignored in both scoring and branching; a question mark response resulted in the administration of the next item in the stratum.

This modified stradaptive test was administered to a group of 113 subjects, and 79 of them were retested. Results, when compared to those of the original version of the stradaptive test, showed substantial reductions in the utility of the consistency scores for predicting stability. The clusters of ability scores and the ranking of those scores with respect to stability were the same as found in the original stradaptive test. Analysis of the question mark responses showed that subjects omitted items that were more difficult than those which they answered incorrectly. It was concluded, therefore, that question mark responses in the stradaptive test should be treated as incorrect responses, rather than being ignored in branching decisions and in scoring.

Three suggestions were made for future research on the stradaptive test. They were 1) monte carlo investigation of the utility of the initial ability estimates; 2) monte carlo investigation of the utility of different termination criteria; and 3) development of improved scoring methods.

The other study of the stradaptive strategy (Waters, 1974, 1975) investigated the modified version of the test in which omissions were ignored. Tests in this study were also administered by CRTs connected to a time-shared computer. The design allowed the calculation of both parallel forms reliability and validity coefficients. Validity was operationalized as the correlation of scores obtained from the stradaptive test with scores from a conventional test composed of similar items taken earlier. Waters' study did not include any analyses of the consistency scores.

The major findings of Waters' study were 1) that the stradaptive strategy was able to attain parallel forms reliabilities and validities comparable to a conventional test having twice as many items; 2) that the relative quality of the scores with respect to reliability and validity was strongly dependent on the termination criterion used; and 3) that the average difficulty of all items answered correctly was consistently one of the best ability level scores in terms of parallel forms reliability and validity.

Waters' suggestions for future research were that future studies concentrate on criteria for test termination and on three ability scores--the inter-

polated stratum difficulty, the average difficulty of all items answered correctly, and the average difficulty of all items correct at the most difficult stratum with greater than chance responding.

From the two empirical studies done on the stradaptive testing strategy, two definite findings have emerged. First, of the scoring methods used, the average difficulty of all items answered correctly has consistently been the best with respect to test-retest stability, parallel forms reliability and correlation with an external criterion. Second, all scores have higher stabilities when omitted items are counted as incorrect responses. In addition, the stradaptive strategy was found to yield higher alternate forms reliability and validity than the conventional test, in one study, and there were some data suggesting that the consistency scores were predictive of retest stability.

Computer Simulation as a Supplement to Live-Testing Studies

While empirical studies do yield important findings--indeed the stability of scores and the effect of ignoring omitted items could only have been evaluated in an empirical study--there are some questions that these studies are ill-equipped to answer. Perhaps the most obvious shortcoming of an empirical study is the fact that it takes a considerable amount of time to test a large number of real subjects. Furthermore, real subjects require real items, and it is very difficult to obtain real items with appropriate characteristics to evaluate some questions of interest, such as, which strategy is better--conventional or adaptive? In addition, too few live subjects have appropriate abilities for evaluating tests at extremely high or low ability levels.

But the most restrictive shortcoming of live-testing studies is the fact that the testee's ability level remains unknown to the psychometrician. This fact precludes calculation of various indices of how well the ability estimate derived from the test reflects the testee's true ability level. One important index of the goodness of a testing procedure, which uses both estimated and true ability, is Birnbaum's (1968) information function. The information function adequately reflects the adaptive test's goal of equiprecision of measurement; equiprecision implies that scores at all ability levels will reflect true ability with the same degree of precision. On the other hand, correlation coefficients and reliability indices, which are generally available from empirical studies, are weighted by the distributional characteristics of the trait within the examinee group (Sympson, 1975), and therefore do not provide data from which equiprecision of measurement can be determined.

Given a question which an empirical study is not equipped to investigate, there are two alternative research approaches--theoretical studies and monte carlo stimulation studies. The theoretical study evaluates a strategy by varying parameters of interest on a purely mathematical basis. It is conceptually superior to the monte carlo simulation study because it eliminates error. But, due to the complexity of some testing strategies (e.g., the stradaptive strategy), simplifying assumptions (e.g., perfectly peaked tests or subtests) are necessary which limit the generalizability of the theoretical study to the world of real data and people. Furthermore, in a strategy like stradaptive, the number of calculations necessary in a theoretical study is prohibitive. Like the theoretical study, the simulation study is based on a mathematical model rather than live

subjects. But rather than calculating exact test characteristics, they are estimated via a stochastic process.

Simulation studies are usually run on computers. The computer first simulates a testee (which usually consists of randomly generating an ability level) and then it generates a sequence of item responses on the basis of a mathematical model of how a real subject with the given ability level would respond to that set of items. Thousands of "subjects" are usually run and the data are analyzed as real data would be. However, since ability levels are known, information curves and other statistics which require both ability levels and ability estimates can be calculated. Like live-testing studies and unlike theoretical studies, summary statistics on simulation study data are subject to random fluctuations. But this is not usually a problem if large numbers of testees are simulated.

The study reported herein is a simulation study of the stradaptive testing strategy. In this study, findings from live-testing studies were re-examined using the simulation technique, and some questions not answerable through live-testing studies were investigated.

METHOD

Design

A simulation study is valuable only to the extent that the underlying model accurately reflects data from live-testing studies. For this reason, the initial phase of this study entailed a simulated replication of the empirical study by Vale & Weiss (1975). Exact replication was not possible, however, since the empirical study used test-retest stability as an evaluative criterion. However, test-retest correlations in a simulation study are, strictly speaking, parallel forms reliability coefficients (Betz & Weiss, 1975) and are formally related to the correlation between test scores and the "true" (i.e., generating) ability. The latter correlation is equivalent to the index of reliability, and the square of that correlation is equivalent to a parallel forms reliability coefficient. As in the live-testing study, intercorrelations among the scores were also calculated. Further analyses not possible in the empirical study, such as calculation of information functions, were also done on the original version of the stradaptive testing strategy used in the live-testing study (here referred to as Variable-Length Stradaptive).

The major aim of the present study was a comparative analysis of the characteristics of stradaptive test scores and conventional test scores under varying conditions. The characteristics of greatest interest were 1) intercorrelations among the scores; 2) correlations between generating ability and the scores; and 3) information provided about ability by the scores at various levels of ability.

The conditions varied within the stradaptive test were 1) the scoring method; 2) the discriminating power of the items; 3) the quality of prior information available about ability; and 4) the number of items administered. The discriminating power was varied by using one of three hypothetical item pools, described below, with item discrimination fixed at a=0.5, a=1.0, or a=2.0. For the replication of empirical results using Variable-Length

Stradaptive, a fourth pool containing parameters of real items with varying discrimination was also used. The quality of prior information was varied by providing the strategy with an initial ability estimate either fixed at Θ=0.0, or distributed normally with a mean of zero and a standard deviation of 1.0, and correlating .0, .5, or 1.0 with generating ability. The extreme correlations were chosen to provide the upper and lower bounds on the effect of initial ability estimates and the .5 correlation was chosen as a typical value. The fixed test lengths investigated in this study were 10, 20, 40, and 60 items.

In order to allow several levels of all conditions to be completely crossed, a simplified version of the stradaptive strategy (referred to as Fixed-Length Stradaptive) was adopted. Fixed-Length Stradaptive had a simpler administration strategy, used fewer scoring methods, and had a fixed termination criterion (thus allowing test length to be manipulated). One further analysis, outside of the crossed design, was done on Fixed-Length Stradaptive. Three potential termination criteria were evaluated on the basis of how well they correlated with error of measurement. This was done to determine whether flexible termination of the stradaptive test would be useful in providing equiprecise measurement.

Tests

Conventional Test

A conventional test was included for purposes of comparison. Items in this test were simply administered in a linear order (i.e., with no branching on the basis of responses) and the test was scored by calculating the proportion of items answered correctly.

Variable-Length Stradaptive

The logic. This test was identical to the test used by Vale & Weiss (1975) except for some minor changes in scoring strategies. On the basis of an initial ability estimate $\hat{\Theta}_I$, the "testee" was given the first item in one of the nine available strata. The stratum, S, from which the first item was administered was determined by rounding the function $S = \hat{\Theta}_I/.65 + 5$ when $1 \leq S \leq 9$, and set to the nearest end point when outside that interval.

If the testee's response to the first item was correct, he was branched to ("administered" an item from) the next more difficult stratum. If his response was incorrect, he was branched to the next easier stratum. If there was not a sufficiently easy or difficult stratum for the required branching, (i.e., when an incorrect response was given to an item in the least difficult stratum, or a correct response to an item in the most difficult stratum) the testee was given another item in the same stratum. Testing continued until a termination criterion was reached. The termination criterion used for Variable-Length Stradaptive was the identification of a stratum in which 20% or less of the items were answered correctly after at least five items had been administered.

The scores. Fifteen scores--ten ability scores and five consistency scores--were evaluated by Vale & Weiss (1975) and are described in detail in that report. These scores, which were also examined in this study, were as follows:

Ability Scores

Score 1. The difficulty of the most difficult item answered correctly.

Score 2. The difficulty of the (N+1)*th* item or the next item that would have been administered had testing continued.

Score 3. The difficulty of the most difficult item answered correctly at a stratum less difficult than the ceiling stratum (i.e., the most difficult item in the most difficult stratum having a chance proportion or less correct); or, if no real item existed, the difficulty of a hypothetical item (i.e., the average difficulty of items that would be in the hypothetical stratum if it existed) in a hypothetical stratum below the lowest available stratum.

Scores 4, 5, and 6. The average difficulties of all items in the strata in which items determining scores 1, 2 and 3 respectively are found.

Score 7. The interpolated stratum difficulty, mathematically defined as:

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

where $\bar{D}_{c-1}$ is the average difficulty of the items in the (C-1)*th* stratum, and C is the ceiling stratum. P_{c-1} is the proportion correct at the (C-1)*th* stratum and S is $\bar{D}_c - \bar{D}_{c-1}$ if $P_{(c-1)} \geq .50$ or $\bar{D}_{c-1} - \bar{D}_{c-2}$ if $P_{(c-1)} < .50$

Score 8. The average difficulty of all items answered correctly.

Score 9. The average difficulty of all items answered correctly between the ceiling stratum and the basal stratum (i.e., the most difficult stratum in which all items administered were answered correctly). Defined as $(\bar{D}_c - \bar{D}_{c-1})/2$ if ceiling and basal strata were adjacent.

Score 10. The average difficulty of items answered correctly in the (c-1)*th* stratum, or, the difficulty of the hypothetical stratum immediately below the easiest stratum available if the testee failed to respond correctly at greater than chance rate in any stratum.

Consistency Scores

Score 11. The standard deviation of difficulties of all items administered.

Score 12. The standard deviation of difficulties of all items answered correctly.

Score 13. The standard deviation of difficulties of all items answered correctly between the ceiling and basal strata.

Score 14. The difference in average difficulties of ceiling and basal strata.

Score 15. The number of strata between the ceiling and basal strata.

Fixed-Length Stradaptive

The stradaptive testing strategy was an important addition to the strategies of adaptive testing because it took a realistic account of the practical testing situation--items were structured in an efficient manner, available prior information was used, and a flexible termination rule was implemented. But some of these practical virtues rendered the strategy very difficult to evaluate. In live-testing, the initial ability estimates inflated test-criterion correlations somewhat and the flexible termination made construction of a comparable conventional test difficult. Thus, for research purposes, it was appropriate to develop a simpler version which would be easier to evaluate.

Further changes in the strategy were suggested by previous research and other considerations. The two major changes involved eliminating some of the scoring strategies and ceasing to use the ceiling and basal strata. As was discussed earlier, score 8, the average difficulty of all items answered correctly, was consistently the best of the original ten ability scores in empirical studies. Thus, only that score was used in the analysis of Fixed-Length Stradaptive.

All consistency scores which required finding the ceiling or basal strata were also eliminated from this part of the study. Although conceptually simple, locating these two strata required some rather complex logic for subjects whose ability was at the extreme upper or lower end of the item pool. In these cases, the ceiling and/or basal strategies were essentially arbitrarily determined. Since the simulation study would result in substantial numbers of testees with extreme ability levels, these scores were not used in this study to eliminate the effects of such arbitrary decisions.

The logic. The administration logic of Fixed-Length Stradaptive was identical to that of Variable-Length Stradaptive except for the termination criterion. A testee was given the first item in one of nine strata chosen on the basis of the same function of initial ability estimate used for Variable-Length Stradaptive. Following a correct response he was branched to the next more difficult stratum and following an incorrect response was branched to the next easier stratum. This process continued until a predetermined number of items had been administered (in this study either 10, 20, 40, or 60 items). Flexible termination was not used in this version of the stradaptive strategy.

The scores. Six scores were calculated for Fixed-Length Stradaptive--three ability scores and three scores intended to predict errors of measurement. The

first score was score 8 from Variable-Length Stradaptive--the average difficulty of all items answered correctly. This score was included since it was the best of the scores in the stradaptive live-testing studies.

The second score was the average difficulty of all items administered. This was a modification of the Variable-Length Stradaptive's score 8, the average difficulty of all items answered correctly, and has been used for scoring other types of adaptive strategies (e.g., Lord, 1970; Larkin & Weiss, 1974; Weiss, 1974). Average difficulty of all items administered was investigated because it is less apt to be affected by erratic response records. For example, consider the case of a fixed termination rule, such as "stop after 40 items." A person might begin the test in the easiest stratum and incorrectly answer the first 31 items. That same person might, by chance, answer the last nine items correctly thus progressing to the most difficult stratum. That person would obtain the same average difficulty correct score as a person who answered 20 items correctly in the fifth stratum and 20 items incorrectly in the sixth stratum. In these two cases, however, the second testee would have encountered more difficult items, on the average, and answered more of them correctly.

The third score was Owen's (1969) Bayesian scoring technique. This scoring method was included as a mathematically "optimal" score, for comparison to the rational scoring methods. Population parameters (i.e., a normal prior ability distribution with mean of zero and standard deviation of 1.0) were used to initialize the scoring procedure for all subjects, regardless of their entry point, and the score was updated after each item response. The FORTRAN IV subroutine used to calculate this score is included in Appendix A.

The three error predictor scores were included in this study for evaluation as termination criteria to be used for flexible termination. The first two were based on the consistency scores investigated by Vale & Weiss (1975). The original consistency scores considered only variability of the response record and not its length. Length has been explicitly taken into consideration in the error predictor scores in a manner analogous to the way that the number of observations is taken into consideration in calculation of the standard error of a mean.

Score 4 is the standard deviation of the difficulties of all items answered correctly divided by the square root of the number of items answered correctly. Score 5 is the standard deviation of the difficulties of all items administered divided by the square root of the number of items administered. Score 6 is the standard error provided by Owen's (1969) formula (i.e., the square root of the Bayesian posterior variance after the last item has been administered).

Item Pools

Obtaining item pools for live-testing studies is a long and tedious process involving writing the items, administering them, norming them, and selecting those most approporiate to the test being constructed. With the relatively large item pools required by adaptive testing strategies, it is difficult to investigate the effects of varying item parameters because a sufficient number of items is generally not available. In simulation studies, however, acquisition of item pools is very easy, since once the desired parameters are specified, the item pool is available. Item pools used in simulation studies do not contain real items with real content but are rather simply a set of parameters of hypothetical items.

Certain limitations, stemming from the mathematical model on which item responses are based, are inherent in simulation studies. Because the items lack content, assumptions made by some test models (such as local independence) are explicitly programmed into the response model and the possibility that these assumptions may be violated by real subjects is ignored. Because of the simplicity of the response models used, effects of memory and thus test-retest stability cannot be examined. Simulation studies are not meant to be a substitute for empirical studies and the simplicity of item pool construction is not without its costs. But in a simulation study, a variety of item pool conditions can be readily constructed. Consequently, some questions which cannot be investigated in live-testing studies, such as the amount of information provided by a testing procedure, can be investigated.

A Real Item Pool

For the simulated replication of empirical findings with Variable-Length Stradaptive, the parameters of the 269 items used for the modified stradaptive test by Vale & Weiss (1975) were used. These parameters, as well as summary statistics, are included in Appendix Table B-1. Although the item parameters used in the present study were those of the modified stradaptive test used in the live-testing study, the branching procedures used were those of the original version.

Hypothetical Item Pools

Although use of parameters obtained from real item pools retains an element of reality not possessed by use of purely hypothetical item pools, it is not feasible to manipulate the item parameters of interest by this procedure. To investigate the effects of item discrimination on test characteristics, three stradaptive test item pools with normal ogive discrimination indices (a) of .5, 1.0 and 2.0 were generated. Since the effects of variability in item discriminations were not investigated in this study, discriminations were held constant within each of the three item pools.

Item difficulty parameters were generated separately for each item pool. These parameters were randomly and rectangularly distributed within each of nine equally spaced strata. Difficulty parameters for each of these three item pools are included in Appendix Tables B-2, B-3 and B-4.

Also generated were three item pools for the conventional test. These pools had constant normal ogive discrimination indices of .5, 1.0 and 2.0 and were randomly and rectangularly distributed within the same range of difficulty as the middle stratum of the stradaptive item pools (i.e., between b=-.33 and b=+.33). Difficulty parameters for these pools are included in Appendix Table B-5.

Generation of the Data

Data in this simulation study were obtained in a way very similar to the way data are collected in a live-testing study. A testing strategy program chose an item, administered that item, and on the basis of responses to several items, calculated a score. The difference between this study and a live-testing study was that in an empirical study, each item is administered to an actual testee, while in this study items were "administered" to an item response

simulator. The simulator then assessed the item parameters and a testee's generated ability level and generated a "correct" or "incorrect" response.

The response simulator used a two-step procedure. First, the probability of a correct response given the "testee's" ability was calculated from the following equation suggested by Lord (1970):

$$P_i(\Theta)=C_i+(1-C_i)\Phi[a_i(\Theta-b_i)] \qquad [1]$$

where $P_i(\Theta)$ ≡ probability that a testee with ability Θ will correctly answer an item,
C_i ≡ probability of a correct answer due to guessing (set to .20 for this study),
$\Phi[x]$ ≡ the unit normal distribution integrated from $-\infty$ to the standard deviate, x,
a_i ≡ the discriminating power of the item,
Θ ≡ the ability level of the testee,
b_i ≡ the difficulty of the item.

After the probability of a correct response was determined, a random number was generated from a rectangular distribution between 0 and 1. If this random number was greater than the probability of answering the item correctly, the item was considered answered incorrectly; otherwise it was considered correct. This procedure is identical to that used by Betz & Weiss (1974), and has been used in a variety of other studies in a slightly different form (e.g., Jensema, 1972; Urry, 1970, 1974).

Generation of the underlying ability was done in two different ways; this is discussed below. The computer and computer programs used in this study are described in Appendix C.

Data Analysis

Descriptive Statistics

Means and standard deviations were calculated for all scores of Variable-Length Stradaptive under all four variations of initial ability estimates using each of the four item pools. These statistics were, in each condition, computed from administration of the test to 1000 hypothetical testees with abilities sampled from a normal distribution with mean of zero and standard deviation of one. These statistics were computed on Variable-Length Stradaptive primarily for comparison with empirical data obtained previously.

Means and standard deviations were calculated for all scores of Fixed-Length Stradaptive to determine how the new scores were distributed, and to assess the effects of varying characteristics of the test. These statistics were calculated for the four lengths of 10, 20, 40 and 60 items under all conditions of initial ability estimates, but using only item parameters of the three hypothetical pools. As with Variable-Length Stradaptive, these statistics were computed from administration of the test to 1000 hypothetical testees with abilities sampled from a normal distribution with mean of zero and standard deviation of 1.0.

Correlational Analysis

Inter-score correlations. Correlations among scores were calculated for both Variable-Length Stradaptive and Fixed-Length Stradaptive. These inter-correlations were calculated using the initial ability estimate fixed at $\theta=0$, on 15,000 hypothetical testees sampled from a normally distributed population.

Score-ability correlations. In classical test theory (Gulliksen, 1950), the correlation of ability and test score is referred to as the "index of reliability". In modern test theory (Lord & Novick, 1968; Urry, 1970) it is referred to as "validity". In live-testing research it is estimated by taking the square root of an alternate form reliability coefficient. In simulation research, it is calculated directly since "ability" is known. Score-ability correlations are useful if the researcher is interested in assessing how well test scores predict ability for some specified population as a whole.

These correlations were calculated for all tests under all variations of conditions. Within each of the conditions, this correlation was calculated on the same sample of 1000 testees used for the descriptive statistics.

In addition to providing a comparison among scores in the simulation study, the squared index of reliability for Variable-Length Stradaptive should be comparable to a parallel forms reliability coefficient, such as the one reported by Waters (1974, 1975). Because of the effect of memory in a test-retest design, it should be somewhat less directly comparable to test-retest stabilities such as those reported by Vale & Weiss (1975).

Information Analyses

While the correlation between test score and ability is a relevant index of how well a score predicts ability for a whole population, it provides little information about how a score predicts ability level within different levels of that ability. For example, a score-ability correlation for a conventional test may be higher than a score-ability correlation for an adaptive test, even though the adaptive test provides a higher precision of measurement at the extremes of ability, simply because correlations are strongly influenced by the larger number of observations in the middle range of a normally distributed population. Since adaptive tests distribute their precision throughout the range of ability, while precise measurement for peaked conventional tests is concentrated in the middle range of ability, a score-ability correlation is a statistic biased in favor of the conventional test, and is not an optimal statistic for comparison of the two strategies. Thus, evaluation of two testing strategies in terms of other criteria, which are less influenced by the distribution of ability in the population, is desirable (see Sympson, 1975, for a discussion of evaluation criteria).

Information curves. The information provided by a score about an ability at some level of that ability is roughly analogous to the precision of measurement at that point, or the ability to discriminate between two arbitrarily close ability levels centered on that point (Lord, 1970). The graph of these information values plotted against all values of ability is called the information curve. In this study, as in Betz & Weiss (1974), information curves

were constructed by calculating information values from a formula, suggested by Birnbaum (1968), at several points along the ability continuum.

Birnbaum's formula is:

$$I_x(\theta) = \left[\frac{\frac{\partial}{\partial\theta} E(X|\theta)}{\sigma_{X|\theta}}\right]^2 \qquad [2]$$

where $I_x(\Theta)$ is the information about Θ provided by score x. The numerator of Equation 2 may be viewed as a transforming function, converting the score, x, into ability units. It is also the partial derivative of the score with respect to ability (Θ) evaluated at a particular level of ability, indicating the relative rate of change of the two variables. The denominator is simply the conditional standard deviation of the score, or the dispersion of the score, x, evaluated at a fixed level of ability (i.e., imprecision of measurement).

To calculate information values, 1000 response records were generated at each of fifteen equally spaced levels of ability ranging from -3.5 through 0.0 to +3.5. For the middle thirteen points, partial derivatives of the score means were calculated with respect to ability at each level of generating ability by taking the derivative of the second degree Lagrangian interpolation polynomial fitted to three sucessive points. This technique finds the first derivative of the second degree polynomial best fitting the point of interest and the two adjacent points. Because points on each side of the point of interest were needed to estimate the polynomial, the endpoints (i.e., -3.5 and +3.5) were not considered in calculating the information values. When the derivatives were obtained, they were divided by the standard deviation of the scores at the level of ability on which the derivative was centered and then squared to yield the information at that point. Connecting the thirteen values of information thus calculated yielded an information curve.

Statistics descriptive of information curves. Graphs are a simple and sufficient way to present information curves, but they are difficult to compare when many information curves are involved. Thus, for economy of presentation, the means and coefficients of variation of information values at the thirteen points defining the information curves were computed.

The mean or average information was computed for each of the information curves. Mean information is a statistic that is not disproportionately weighted by ability distribution characteristics as is the correlation coefficient. The higher the mean information, the more information the score provides about ability at all levels of ability, on the average. Higher mean information implies better measurement.

The coefficient of variation was also computed for each information curve. This statistic is of interest because its departure from zero means that the goal of equiprecision of measurement is not being achieved. This index is equal to the standard deviation of the thirteen points of the information curve divided by their mean and multiplied by 100 (see Guilford, 1950, p. 118,

for a discussion of the coefficient of variation). It was chosen over the standard deviation because, unlike the standard deviation, it is not affected by the absolute magnitude of the information curve. For example, the height of the information curve of a conventional test composed of equivalent items is directly proportional to the length of the test and therefore, the mean and standard deviation of the values of the information curve are similarly proportional to length. Double the length of a conventional test and the mean and standard deviation of the information curve both double. The coefficient of variation, being doubled in both the numerator and the denominator would remain constant, however.

The choice of the coefficient of variation as an evaluative criterion involved the value judgment that the relative rather than the absolute variation was important in evaluating the equiprecision provided by a score on a test. The information curve of an ideal test--one with measurement of equal precision throughout the ability range--would have a high mean and a coefficient of variation equal to zero.

Termination Criterion Analysis

One goal in the design of the stradaptive strategy was to identify response records in which the test was not unambiguously locating the testee's ability level, so that the test could be extended in length to provide better precision of measurement for that testee. Thus, it was intended that a flexible termination criterion be used so that different individuals could be administered different numbers of items. In this study, the three error predictor scores of Fixed-Length Stradaptive (scores 4, 5 and 6) were examined as possible candidates for the termination criterion.

To perform these evaluations, 1000 administrations using each of the three hypothetical stradaptive item pools were randomly terminated after an average of 30 items; the standard deviation of number of items at termination was 6. The tests were terminated at various lengths because a termination criterion must be effective at all lengths, since a real test might terminate at any of several lengths. The error scores were not allowed to function as termination criteria in this analysis because an effective termination criterion would hold the error of measurement constant. Then, with no variability in error of measurement, there could be no correlation between it and test score.

To provide a criterion of error of measurement to predict, ability scores and generating ability were first standardized within each set of 1000 administrations to account for differences in scoring metric. Then the unsigned differences between the standardized ability scores and the standardized generating ability were calculated. Error predictor scores were then correlated with these absolute errors, since it was expected that a useful termination criterion would correlate with these errors. It may be noted that this procedure is identical to Ghiselli's (1956) procedure for discovery of moderator variables. The search is for a variable to correlate with absolute deviations from the line of relations.

RESULTS

Analysis of the Conventional Test

Descriptive Statistics

The means and standard deviations of the conventional test scores with varying test lengths and item discriminations are presented in columns three and four of Table 1. The mean proportions correct were all close to a value of .60 which was only slightly higher than the value of .588 obtained by Vale & Weiss (1975) in their live administration of a 40-item conventional test. A slight difference in the opposite direction was expected because the test used in the empirical administration had easier items ($\bar{b}$=-.368).

Table 1

Summary Statistics for the Conventional Test as a Function of Item Discrimination and Test Length

Discrimination (a)	No. Items	Mean	S.D.	Correlation with Ability	Information	
					Mean	Coefficient of Variation
0.5	10	.612	.209	.703	.725	42.046
	20	.607	.179	.811	1.448	40.290
	40	.598	.162	.887	2.882	41.232
	60	.600	.158	.917	4.307	39.617
1.0	10	.616	.267	.851	1.771	75.451
	20	.592	.250	.908	3.198	88.171
	40	.600	.243	.938	6.444	87.808
	60	.597	.238	.950	9.595	87.978
2.0	10	.597	.326	.888	3.484	138.147
	20	.592	.317	.906	6.601	139.150
	40	.605	.311	.918	13.630	133.632
	60	.612	.307	.926	19.674	135.361

This discrepency may have been due to the fact that the items used in the live-testing study were normed on relatively small groups of subjects (see McBride & Weiss, 1974). Waters' (1974, 1975) conventional tests had a mean proportion correct of .665, which was expected because his items were easier ($\bar{b}$=-.368) and normed on larger subject groups. No trend among the score means across varying test lengths and item discrimination was apparent.

Standard deviations of conventional test scores showed two trends. As item discriminating power increased, the standard deviations of the scores increased. As test length was increased, standard deviation decreased.

Correlational Analyses

Column five of Table 1 presents the correlations of conventional test scores with generating ability. The results show, as expected, that the correlation increases as the length of the test is increased. A more complex trend was observed with respect to the item discrimination. The score-ability correlation increased with increasing item discrimination, and then tapered off. The correlation between score and ability was a joint function of test length and item discrimination. For the 10-item test the score-ability correlation increased with increases in item discrimination. For the 20-item test, the score-ability correlation improved as discriminations were increased from .5 to 1.0, but remained about the same as discriminations were increased to 2.0. On 40- and 60-item tests, score-ability correlations improved when discriminations were increased from .5 to 1.0, but were lower for items of 2.0 discrimination.

The attenuation paradox (Loevinger 1954; Sitgreaves, 1961) is the apparent reason for the lower score-ability correlations with higher item discriminations. The attenuation paradox refers to the fact that as items get more discriminating, they provide more information at a point and less information at abilities distant from that point. The conventional test had items all of similar difficulty, and as items became more discriminating, it measured less accurately for testees with abilities outside of an increasingly narrow range.

Information Analyses

Column six of Table 1 shows the average information provided by the conventional test. Average information increased in almost direct proportion to test length, a result that was expected from modern test theory (Lord & Novick, 1968). It also increased in almost direct proportion to item discriminating power. The decrease observed in score-ability correlations as item discriminations became high (2.0) was not observed with respect to average information.

Column seven of Table 1 presents the coefficient of variation, an index of equiprecision of measurement. With discriminations held constant, this index remained relatively constant across tests of different length, as it was expected to do. It increased considerably, however, with changing item discriminations, indicating that the conventional test provides relatively poorer equiprecision of measurement as items become more discriminating. The trend involved seemed to indicate that the coefficient of variation is directly proportional to the item discriminations, doubling when discriminations are doubled. The relation was not completely linear, however, as the coefficient of variation with discrimination of 2.0 was less than the value of about 170 that would be anticipated from a strictly linear relationship.

Analysis of Variable-Length Stradaptive

Descriptive Statistics

Table 2 presents the means of the fifteen Variable-Length Stradaptive test scores (scores 1-10 are ability scores, and 11-15 are consistency scores). Independent variables are 1) the initial ability estimate, fixed at 0.0 and randomly distributed N(0,1) correlating 0.0, .5 and 1.0 with the generating

Table 2
Mean Scores for Variable-Length Stradaptive
as a Joint Function of Ability Estimate Validities,
Item Discriminations (a)
and Average Number of Items Administered

Score	a	Fixed Entry	Initial Ability Correlation 0.0	0.5	1.0	Score	a	Fixed Entry	Initial Ability Correlation 0.0	0.5	1.0
1	0.5	1.580	1.660	1.600	1.540	9	0.5	.208	.201	.162	.200
	1.0	1.200	1.220	1.170	1.170		1.0	.053	.033	.043	.013
	2.0	.888	.915	.880	.802		2.0	-.048	-.037	-.055	-.062
	variable	1.350	1.430	1.320	1.300		variable	.202	.226	.194	.173
2	0.5	.840	.811	.793	.818	10	0.5	.813	.821	.811	.824
	1.0	.565	.542	.573	.523		1.0	.494	.456	.471	.451
	2.0	.276	.310	.271	.253		2.0	.163	.181	.178	.135
	variable	.770	.782	.761	.782		variable	.665	.699	.655	.643
3	0.5	1.020	1.030	1.020	1.030	11	0.5	.834	.875	.862	.814
	1.0	.732	.691	.707	.691		1.0	.743	.785	.742	.725
	2.0	.385	.400	.392	.361		2.0	.673	.713	.672	.625
	variable	.875	.908	.869	.858		variable	.757	.813	.752	.724
4	0.5	1.490	1.560	1.500	1.450	12	0.5	.787	.827	.815	.760
	1.0	1.130	1.180	1.090	1.100		1.0	.669	.713	.670	.652
	2.0	.803	.834	.791	.713		2.0	.575	.607	.575	.529
	variable	1.270	1.360	1.240	1.230		variable	.698	.755	.693	.663
5	0.5	.839	.819	.801	.823	13	0.5	.498	.508	.522	.499
	1.0	.559	.541	.569	.515		1.0	.401	.396	.398	.404
	2.0	.292	.317	.279	.269		2.0	.268	.272	.275	.271
	variable	.793	.802	.785	.803		variable	.405	.424	.414	.417
6	0.5	.829	.837	.824	.842	14	0.5	2.550	2.590	2.650	2.580
	1.0	.522	.480	.497	.476		1.0	2.110	2.080	2.080	2.120
	2.0	.181	.196	.189	.155		2.0	1.620	1.620	1.630	1.610
	variable	.695	.729	.685	.676		variable	2.140	2.190	2.150	2.170
7	0.5	.887	.887	.878	.900	15	0.5	2.910	2.970	3.050	2.950
	1.0	.577	.537	.550	.534		1.0	2.230	2.180	2.190	2.240
	2.0	.262	.273	.269	.236		2.0	1.480	1.480	1.490	1.460
	variable	.760	.794	.747	.734		variable	2.280	2.360	2.300	2.320
8	0.5	.181	.192	.156	.167	Test	0.5	54.000	56.100	55.800	55.100
	1.0	-.006	-.027	-.019	-.025	Length	1.0	40.800	39.600	39.700	40.200
	2.0	-.116	-.131	-.129	-.126		2.0	28.400	29.400	29.500	27.100
	variable	.126	.127	.108	.114		variable	31.900	32.000	31.200	31.800

ability; and 2) three item pools with discriminations fixed at a=.5, 1.0 and 2.0, and the real item pool with varying discriminations that was previously used in an empirical study of the stradaptive test (Vale & Weiss, 1975). Table 2 shows two trends in the means as a function of item parameters and initial ability estimates. The means of all fifteen scores became lower as the items became more discriminating, and the means of the consistency scores, scores 11 to 15, decreased with increasingly valid initial ability estimates.

The trend in the means of the ten ability scores to decrease as a function of item discriminations is at least partly due to the effects of guessing. Implicit in the up-one, down-one branching strategy used in the stradaptive test is the goal of converging on items of difficulty such that the testee's probability of answering correctly is .5. Since all ten Variable-Length Stradaptive ability scores are some rough monotonic function of this difficulty, anything that affects that difficulty will monotonically affect the ability scores.

Difficulty in terms of probability of a correct response, discrimination, and guessing parameters is obtained by rearranging equation 1:

$$b=\Theta - \left[\Phi^{-1} \left(\frac{P-C}{1-C} \right) \right] \Big/ a \qquad [3]$$

where $\Phi^{-1}[x]$ is the inverse function of $\Phi[x]$ yielding the standardized normal deviate when given the cumulative proportion.

The difficulty yielding a probability of being correct of .5, assuming a guessing probability of .2, is:

$$b=\Theta + \frac{.32}{a} \qquad [4]$$

in which b is a decreasing function of a. This shows that the optimal difficulty, and thus the ability scores, should decrease as a function of discrimination when guessing is possible (specifically with a probability of .2). When guessing is not possible, equation 3 reduces to:

$$b=\Theta \qquad [5]$$

in which b is not a function of a. Therefore, the ability scores are expected to decrease as a function of item discriminations through their relation with this "optimal" difficulty level only when guessing is possible.

In addition, some scores (e.g., the difficulty of the most difficult item correct) would be expected to decrease as a function of increasing item discrimination even when guessing is not possible due to their joint dependence on the optimal difficulty level and variability of the response record. For example, a testee with an inconsistent response record (i.e., one which ranges across a large number of strata) would be expected to have a more difficult item correct than a testee having a more consistent response record and the same average difficulty of items, simply because he encountered a few more difficult items (along with, of course, a few less difficult items).

The consistency scores were expected to decrease as fewer inappropriate items were administered in the process of locating the appropriate strata from which to administer items (i.e., as the initial ability estimates improved and the entry point became closer to the testee's true ability) and as fewer incorrect branchings were made (due to more discriminating items). An observation worthy of note with respect to the decrease in mean consistency scores as a function of initial ability estimates is that some consistency scores, under some conditions, were not lower using valid initial ability estimates than they were using a fixed entry point at the middle stratum (e.g., the between ceiling and basal strata variability scores, scores 13, 14 or 15) or required a reasonably good initial ability estimate before they improved (e.g., the overall variability scores, 11 and 12).

Mean scores generated using the real item pool were slightly higher in this study than in the empirical study by Vale & Weiss (1975) using nearly the same pool on real subjects (e.g., 1.350 vs 1.073 for score 1; .770 vs .560 for score 2). This may have been due either to inadequacies in the simulation model or inaccurate item parameters in the live-testing study.

Test lengths showed a marked decreasing trend as item discriminations improved, but no apparent trend at all with respect to goodness of initial ability estimates. The average number of items administered using the real item pool varied between 31.2 and 32.0 items; these means were between those obtained using the hypothetical pools with discriminations of a=1.0 and a=2.0. This was not expected because the mean discrimination of the real pool was a=.717. This was an underestimate of the discriminations of the items actually administered, however, because the most discriminating items were placed first in the pool; the corrected average discrimination value in the live-testing study was a=.879. It appears that putting a few highly discriminating items at the beginning of each stratum may drastically shorten the stradaptive test when the original termination criterion is used. The mean length of the identical strategy in the live-testing study using a slightly different item pool (i.e., one with a few less items) varied between 27.8 and 31.4 items. Thus, the mean lengths obtained in this simulation were reasonably close to those obtained with real subjects.

Table 3 shows standard deviations for the 15 scores on Variable-Length Stradaptive for combinations of the same independent variables. For both the ability level and consistency scores there was no apparent trend with respect to initial ability estimates, and only a very slight tendency for the standard deviations to decline as item discriminations improved. Score 8, the average difficulty of all items answered correctly, was the least variable of the ability scores. Scores 11, 12 and 13, the standard deviation consistency scores, had low variability both with respect to the ability scores and the distance variability scores, 14 and 15. The differences among the standard deviations observed in this study were also apparent in the live-testing study, although in this study the values were slightly lower. This latter result was expected, though, since the live-testing study included sources of error not included in a simulation study.

Correlational Analyses

Table 4 presents the intercorrelations among the Variable-Length Stradaptive

Table 3
Standard Deviations of Variable-Length Stradaptive Scores
as a Joint Function of Ability Estimate Validities
and Item Discriminations (a)

Score	a	Fixed Entry	Initial Ability Correlation 0.0	0.5	1.0	Score	a	Fixed Entry	Initial Ability Correlation 0.0	0.5	1.0
1	0.5	1.110	1.100	1.120	1.170	9	0.5	.993	1.050	1.040	1.060
	1.0	1.140	1.080	1.140	1.150		1.0	.982	1.000	1.040	1.030
	2.0	1.080	1.110	1.130	1.130		2.0	1.000	1.000	1.020	1.010
	variable	1.240	1.200	1.270	1.270		variable	1.090	1.090	1.100	1.110
2	0.5	1.180	1.220	1.220	1.200	10	0.5	1.230	1.290	1.280	1.260
	1.0	1.140	1.140	1.150	1.110		1.0	1.170	1.160	1.180	1.130
	2.0	1.180	1.150	1.150	1.150		2.0	1.110	1.110	1.120	1.090
	variable	1.350	1.340	1.290	1.320		variable	1.330	1.310	1.320	1.290
3	0.5	1.260	1.320	1.300	1.290	11	0.5	.201	.216	.215	.222
	1.0	1.190	1.180	1.200	1.150		1.0	.144	.179	.158	.154
	2.0	1.130	1.120	1.140	1.100		2.0	.143	.170	.149	.133
	variable	1.390	1.360	1.370	1.330		variable	.170	.196	.187	.181
4	0.5	1.060	1.040	1.060	1.130	12	0.5	.238	.256	.245	.253
	1.0	1.100	1.050	1.090	1.110		1.0	.182	.217	.190	.184
	2.0	1.060	1.070	1.110	1.110		2.0	.189	.212	.192	.167
	variable	1.170	1.140	1.200	1.220		variable	.221	.247	.232	.215
5	0.5	1.160	1.200	1.190	1.170	13	0.5	.315	.313	.316	.316
	1.0	1.120	1.120	1.130	1.100		1.0	.235	.236	.233	.231
	2.0	1.140	1.110	1.100	1.100		2.0	.186	.184	.193	.172
	variable	1.320	1.310	1.260	1.290		variable	.274	.271	.269	.257
6	0.5	1.220	1.280	1.270	1.260	14	0.5	1.280	1.260	1.290	1.300
	1.0	1.150	1.150	1.150	1.110		1.0	.920	.918	.905	.913
	2.0	1.090	1.090	1.100	1.060		2.0	.650	.649	.694	.598
	variable	1.330	1.310	1.320	1.290		variable	.987	.973	.970	.927
7	0.5	1.180	1.240	1.210	1.220	15	0.5	1.960	1.930	1.970	2.000
	1.0	1.100	1.100	1.120	1.080		1.0	1.410	1.400	1.390	1.400
	2.0	1.050	1.050	1.060	1.040		2.0	.996	.991	1.060	.917
	variable	1.280	1.270	1.280	1.260		variable	1.520	1.500	1.490	1.430
8	0.5	.888	.918	.944	.992						
	1.0	.910	.915	.977	1.010						
	2.0	.917	.933	.976	1.010						
	variable	.961	.971	1.030	1.070						

Table 4
Intercorrelations of Variable-Length Stradaptive Scores

	Ability Scores										Consistency Scores				
Score	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1.000	.806	.893	.979	.818	.894	.892	.889	.839	.893	.397	.544	.423	.424	.425
2	.806	1.000	.881	.786	.986	.883	.883	.838	.861	.879	.114	.240	.402	.372	.373
3	.893	.881	1.000	.870	.890	.996	.991	.906	.921	.997	.193	.296	.507	.497	.499
4	.979	.786	.870	1.000	.795	.875	.875	.877	.826	.872	.430	.541	.400	.402	.402
5	.818	.986	.890	.795	1.000	.892	.892	.845	.869	.888	.112	.245	.406	.379	.379
6	.894	.883	.996	.875	.892	1.000	.994	.904	.920	.997	.199	.295	.502	.505	.507
7	.892	.883	.991	.875	.892	.994	1.000	.929	.943	.990	.173	.270	.437	.435	.437
8	.889	.838	.906	.877	.845	.904	.929	1.000	.979	.903	.073	.203	.178	.156	.158
9	.839	.861	.921	.826	.869	.920	.943	.979	1.000	.919	.006	.122	.182	.152	.154
10	.893	.879	.997	.872	.888	.997	.990	.903	.919	1.000	.198	.301	.514	.508	.511
11	.397	.114	.193	.430	.112	.199	.173	.073	.006	.198	1.000	.920	.560	.545	.540
12	.544	.240	.296	.541	.245	.295	.270	.203	.122	.301	.920	1.000	.579	.544	.541
13	.423	.402	.507	.400	.406	.502	.437	.178	.182	.514	.560	.579	1.000	.951	.953
14	.424	.372	.497	.402	.379	.505	.435	.156	.152	.508	.545	.544	.951	1.000	1.000
15	.425	.373	.499	.402	.379	.507	.437	.158	.154	.511	.540	.541	.953	1.000	1.000

scores. The same four clusters of ability scores observed in the live-testing study were apparent here. The three item difficulty scores, scores 1, 2 and 3, formed three two-variable clusters, each with their respective stratum difficulty scores, scores 4, 5 and 6. Additionally, 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 fourth ability score cluster was composed of scores 8 and 9, the average difficulty scores.

Clustering among the variability scores was obvious and again consistent with the live-testing data. Scores 11 and 12, the overall variability scores, formed one cluster. Scores 13, 14 and 15, the between ceiling and basal strata consistency scores, formed another.

Table 5 presents the correlations of Variable-Length Stradaptive ability scores with the ability that generated the responses. The expected increasing trend in correlation with improving initial ability estimates was not observed with any regularity across different scores and different parameters although the trend was apparent for scores 8 and 9.

Table 5
Score-Ability Correlations for Stradaptive,
as a Joint Function of Item Discriminations (a)
and Initial Ability Validities

Score	a	Fixed Entry	Initial Ability Correlation 0.0	0.5	1.0	Score	a	Fixed Entry	Initial Ability Correlation 0.0	0.5	1.0
1	0.5	.602	.620	.644	.642	6	0.5	.654	.705	.667	.674
	1.0	.768	.731	.784	.773		1.0	.822	.817	.830	.821
	2.0	.798	.805	.807	.833		2.0	.893	.886	.881	.899
	variable	.763	.734	.761	.766		variable	.787	.798	.792	.790
2	0.5	.630	.682	.633	.662	7	0.5	.678	.728	.689	.699
	1.0	.798	.782	.803	.792		1.0	.847	.845	.854	.851
	2.0	.851	.832	.855	.862		2.0	.918	.913	.912	.924
	variable	.757	.783	.770	.766		variable	.810	.816	.811	.811
3	0.5	.656	.707	.669	.676	8	0.5	.805	.809	.838	.872
	1.0	.821	.820	.827	.820		1.0	.918	.894	.918	.938
	2.0	.894	.887	.879	.899		2.0	.950	.935	.954	.960
	variable	.784	.799	.791	.786		variable	.887	.873	.892	.914
4	0.5	.593	.605	.630	.632	9	0.5	.791	.815	.807	.839
	1.0	.762	.713	.771	.767		1.0	.918	.912	.924	.932
	2.0	.795	.797	.811	.834		2.0	.954	.949	.959	.956
	variable	.752	.724	.754	.761		variable	.887	.888	.891	.906
5	0.5	.648	.693	.648	.674	10	0.5	.652	.705	.666	.677
	1.0	.806	.791	.809	.800		1.0	.818	.815	.826	.816
	2.0	.859	.842	.861	.863		2.0	.891	.884	.877	.899
	variable	.770	.791	.777	.772		variable	.785	.797	.789	.788

A definite increasing trend in score-ability correlation with increasing item discrimination was observed, however. This result suggests that there may be an inadequacy in the termination criterion as it should vary the length of the test to keep constant the accuracy of measurement, and thus the score-ability correlation.

The parameters from the real item pool provided score-ability correlations somewhere between the values provided by the hypothetical pools with discriminations of .5 and 1.0. This result is consistent with their average discrimination value of about .88.

Scores 8 and 9 consistently had the highest correlation with ability, a finding consistent with the fact that these two scores showed the highest stability in the live-testing study by Vale & Weiss (1975) and generally highest validities and reliabilities in the study by Waters (1974, 1975).

Comparison with the conventional test. Comparison of the score-ability correlations of Variable-Length Stradaptive with those of the conventional test (Table 1) is difficult because the flexible termination criterion of Variable-Length Stradaptive yields test lengths not directly comparable to those of the conventional test. A rough comparison showed the stradaptive strategy as better using items with discriminations of 2.0. Using stradaptive score 8, and comparing the fixed entry point administration (the fair comparison since the conventional test cannot utilize prior information), the stradaptive test correlated .950 with an average of 28.4 items while the conventional test correlated only .918 with 40 items and .926 with 60 items. Thus, with highly discriminating items, the 28-item stradaptive test correlated higher with ability than did the 60-item conventional test.

The conventional test achieved higher score-ability correlations than the stradaptive test with less discriminating items. With discriminations of 1.0 the stradaptive test score 8 correlated only .918 after an average of 40.8 items while the conventional test correlated .938 after 40 items. At discriminations of .5, stradaptive score 8 correlated .805 with ability after an average of 54 items and the conventional test correlated .887 after 40 items and .917 after 60 items. Thus, Variable-Length Stradaptive testing strategy was superior to a conventional test, with respect to score ability correlations, only when given very discriminating items.

Information Analyses

Table 6 presents the average information values provided by Variable-Length Stradaptive scores. As with the score ability correlations, an increasing trend in information was observed for all scores as item discriminations improved. The values of average information provided by the items from the real item pool were, as were the correlations, between those for item pools with discriminations of .5 and 1.0. The increasing trend observed in the correlations as a function of validity of initial ability estimates was somewhat more apparent in the average information data than it was in the correlation data (notably for scores 1, 4, and 9). It also appeared in some conditions of most other scores. Score 9, the average difficulty of all items answered correctly between the ceiling and basal strata, provided the highest average level of information. Score 8,

Table 6
Average Information Provided by
Variable-Length Stradaptive Scores
as a Joint Function of Item Discriminations (a)
and Initial Ability Validities

Score	a	Fixed Entry	Initial Ability Correlation 0.0	0.5	1.0	Score	a	Fixed Entry	Initial Ability Correlation 0.0	0.5	1.0
1	0.5	.559	.581	.651	.765	6	0.5	.800	.757	.800	.815
	1.0	1.172	1.208	1.450	1.748		1.0	1.915	1.786	1.971	1.968
	2.0	1.823	1.605	2.174	2.520		2.0	3.713	3.930	4.116	3.682
	variable	.926	.805	1.232	1.635		variable	1.683	1.550	1.840	1.939
2	0.5	.757	.755	.847	.814	7	0.5	.981	.948	1.000	1.011
	1.0	1.439	1.518	1.419	1.718		1.0	2.894	2.675	2.835	2.958
	2.0	2.286	2.988	2.545	2.454		2.0	7.825	8.398	8.934	7.956
	variable	1.302	1.251	1.423	1.488		variable	2.038	1.866	2.286	2.381
3	0.5	.811	.780	.800	.826	8	0.5	1.963	1.749	2.001	2.939
	1.0	1.811	1.723	1.853	1.888		1.0	4.664	3.397	5.521	7.514
	2.0	3.441	3.703	3.917	3.502		2.0	9.868	6.809	12.391	12.303
	variable	1.751	1.608	1.880	1.971		variable	2.782	2.127	3.134	4.066
4	0.5	.518	.507	.649	.787	9	0.5	1.909	1.974	2.221	2.460
	1.0	.954	.858	1.341	1.787		1.0	5.786	5.541	6.024	6.939
	2.0	1.391	1.320	1.658	2.025		2.0	11.864	11.593	13.210	13.617
	variable	.875	.745	1.168	1.583		variable	3.409	3.165	3.976	4.589
5	0.5	.819	.787	.900	.861	10	0.5	.786	.764	.795	.808
	1.0	1.587	1.633	1.556	1.794		1.0	1.884	1.779	1.918	1.967
	2.0	2.604	2.914	2.987	2.797		2.0	3.584	3.781	4.119	3.573
	variable	1.385	1.335	1.494	1.528		variable	1.704	1.560	1.863	1.967

the average difficulty of all items answered correctly, which had correlated highest with ability, also provided relatively high levels of information.

Table 7 contains coefficients of variation of the information values for Variable-Length Stradaptive ability scores. No definite trends with respect to either item discriminations or quality of initial ability estimates were apparent, although in 31 out of 42 comparisons a fixed entry point provided more equiprecise measurement than did a random invalid one (i.e., r=0.0).

Comparison with the conventional test. With discriminations of .5 and the entry point fixed, stradaptive score 9 provided an average information value of 1.909 (Table 6) with an average of 54.0 items (Table 2). With 40 items the conventional test (Table 1) provided a higher average value of 2.882. When discriminations were 1.0, stradaptive score 9 provided 5.786 units of information using 40.8 items on the average. This was still below the average value of 6.444 provided by the conventional test with 40 items. When discriminations were 2.0, stradaptive score 9 provided 11.864 units of information with 28.4 items, a value between the values of 6.601 and 13.630 provided by the conventional test with 20

and 40 items, respectively. When Variable-Length Stradaptive was permitted to function as designed, i.e., with a variable entry point, a moderately valid initial ability estimate (r=.50) resulted in average information of 13.2 (based on an average of 29.5 items) for stradaptive, compared to 13.630 for the longer 40-item conventional test, with discriminations of 2.0.

Table 7
Coefficients of Variation for Information Values of Variable-Length Stradaptive Scores, as a Joint Function of Item Discriminations (a) and Initial Ability Validities

Score	a	Fixed Entry	Initial Ability Correlations 0.0	0.5	1.0	Score	a	Fixed Entry	Initial Ability Correlations 0.0	0.5	1.0
1	0.5	34.93	47.36	33.64	21.10	6	0.5	33.48	34.98	36.05	30.92
	1.0	43.83	59.69	24.17	26.67		1.0	18.80	19.53	34.76	25.21
	2.0	63.69	64.64	49.60	28.23		2.0	23.67	26.34	23.84	20.07
	variable	45.47	55.73	44.76	62.09		variable	64.87	55.82	62.68	67.78
2	0.5	35.22	43.99	39.56	63.74	7	0.5	17.93	24.51	38.02	23.29
	1.0	24.59	32.28	33.34	30.87		1.0	31.61	36.19	26.13	26.77
	2.0	32.82	70.96	17.87	36.72		2.0	79.23	103.99	103.13	79.49
	variable	63.23	62.90	68.47	69.46		variable	61.39	50.92	59.10	65.06
3	0.5	30.35	31.88	32.58	29.43	8	0.5	45.27	52.03	45.85	40.82
	1.0	18.85	18.21	30.92	23.05		1.0	50.56	50.06	52.19	44.36
	2.0	22.36	31.66	34.61	15.50		2.0	57.18	58.62	73.13	41.25
	variable	68.70	56.51	65.66	70.52		variable	52.15	69.13	57.24	59.56
4	0.5	43.93	65.70	37.97	50.71	9	0.5	36.36	50.85	52.99	40.80
	1.0	61.96	56.98	31.18	72.89		1.0	43.37	52.39	46.67	29.81
	2.0	67.98	76.03	41.74	32.94		2.0	57.04	55.19	50.30	44.65
	variable	52.10	65.74	51.76	66.21		variable	43.38	39.10	51.41	57.78
5	0.5	36.00	43.59	39.80	58.68	10	0.5	33.11	33.58	35.20	27.89
	1.0	25.30	35.49	32.47	29.62		1.0	15.01	16.83	30.25	20.90
	2.0	34.06	35.12	37.03	33.35		2.0	25.43	34.55	41.13	20.73
	variable	60.35	61.50	66.30	69.08		variable	69.69	57.81	65.90	69.85

With a fixed entry into the stradaptive test, the most informative score, score 9, was more equiprecise than the conventional test score in all comparable conditions. With discriminations of 2.0 the conventional test score showed a coefficient of variation more than twice as large (see Table 1) as that of Variable-Length Stradaptive's score 9 (Table 7). In fact, given equal item discriminations, only two stradaptive scores, scores 4 and 8, were less equiprecise than scores on a comparable conventional test and then only for poorly discriminating items. In general, as item discriminations increased, the relative equiprecision of the stradaptive test became considerably greater than that of the conventional test.

Table 8
Mean Scores for Fixed-Length Stradaptive as a Joint Function of Item Discriminations (a), Test Length, and Validity of Initial Ability Estimates

a	No. Items	Fixed Entry	Initial Ability Correlation 0.0	0.5	1.0	a	No. Items	Fixed Entry	Initial Ability Correlation 0.0	0.5	1.0
		Score 1						Score 4			
0.5	10	.085	.070	.093	.053	0.5	10	.284	.306	.287	.266
	20	.186	.193	.182	.177		20	.240	.258	.241	.230
	40	.257	.245	.255	.261		40	.192	.197	.192	.187
	60	.282	.272	.283	.283		60	.163	.169	.167	.160
1.0	10	-.072	-.083	-.050	-.041	1.0	10	.272	.294	.270	.246
	20	-.003	-.004	.009	.014		20	.215	.232	.217	.203
	40	.065	.033	.040	.048		40	.164	.171	.164	.157
	60	.058	.044	.056	.056		60	.136	.140	.136	.133
2.0	10	-.118	-.130	-.131	-.092	2.0	10	.250	.279	.251	.219
	20	-.091	-.110	-.105	-.082		20	.189	.202	.188	.175
	40	-.083	-.082	-.074	-.069		40	.138	.144	.137	.131
	60	-.061	-.067	-.058	-.072		60	.116	.119	.114	.110
		Score 2						Score 5			
0.5	10	.369	.374	.375	.325	0.5	10	.231	.245	.232	.215
	20	.465	.484	.461	.440		20	.186	.197	.186	.179
	40	.536	.536	.535	.527		40	.143	.147	.144	.141
	60	.559	.559	.562	.556		60	.121	.125	.123	.120
1.0	10	.261	.263	.277	.253	1.0	10	.225	.240	.222	.205
	20	.331	.346	.335	.319		20	.170	.181	.170	.161
	40	.387	.370	.364	.358		40	.126	.130	.126	.121
	60	.378	.377	.378	.369		60	.104	.106	.104	.102
2.0	10	.250	.263	.237	.233	2.0	10	.211	.233	.212	.189
	20	.266	.270	.255	.251		20	.155	.166	.155	.145
	40	.270	.286	.278	.269		40	.112	.117	.111	.107
	60	.290	.295	.294	.270		60	.094	.095	.092	.089
		Score 3						Score 6			
0.5	10	-.001	-.004	.005	-.045	0.5	10	.734	.739	.741	.740
	20	-.008	.001	-.006	-.027		20	.610	.613	.615	.616
	40	.000	.001	.000	-.003		40	.478	.480	.481	.482
	60	.004	-.006	-.006	-.005		60	.407	.409	.409	.410
1.0	10	-.023	-.027	-.014	-.060	1.0	10	.518	.530	.537	.545
	20	-.005	-.001	-.007	-.028		20	.392	.400	.403	.406
	40	.015	-.001	-.004	-.005		40	.288	.289	.290	.292
	60	.004	-.000	.003	-.011		60	.237	.238	.239	.239
2.0	10	.012	.003	-.030	-.048	2.0	10	.358	.377	.386	.400
	20	.005	.003	-.019	-.033		20	.251	.254	.259	.263
	40	-.008	.004	-.007	-.012		40	.172	.174	.175	.176
	60	.002	.007	.007	-.013		60	.139	.140	.141	.141

Analysis of Fixed-Length Stradaptive

Descriptive Statistics

Table 8 presents the means of Fixed-Length Stradaptive scores as a function of item discrimination, test length, and initial ability estimate. Table 8 shows that, as with Variable-Length Stradaptive, the means of scores 1 and 2, the average difficulty scores, tended to decrease as item discriminating power increased. Score 3, the Bayesian score, showed no such tendency, however. This was probably because the Bayesian score is not directly related to the "optimal" item difficulty discussed previously. Means of scores 1 and 2 tended to increase, as the tests became longer, at all levels of item discrimination. The Bayesian score, score 3, showed no trend with respect to test length. Score 2 means increased as the validity of initial ability estimates increased, but scores 1 and 3 evidenced no such trend.

The corresponding error predictor scores, scores 4, 5 and 6, decreased both as item discriminating power and test length increased. These results were consistent with the desired characteristics of a score designed to predict precision of measurement; errors of measurement of test scores should decrease as both test length and item discriminating power increase.

Means of scores 4 and 5 decreased as the validity of the initial ability estimate improved, but score 6 showed a slight increase. It is not clear whether this was because the initial ability estimate has no advantageous effect on precision of measurement or because the Bayesian error score does not predict errors of measurement.

Table 9 presents the standard deviations of Fixed-Length Stradaptive's scores. The Bayesian ability score, score 3, showed increased variability as item discriminations increased, but no trends with respect to item discriminations were apparent for scores 1 or 2. The trend in score 3 may be because the Bayesian procedure used implicitly regresses scores toward the mean and this regression is less pronounced as items get more discriminating and measurement becomes more precise.

All three ability scores showed increasing variability with increasing test length, when a fixed entry point was used. However, for scores 1 and 2, variability seemed to be a joint function of number of items, item discriminations, and the validity of initial ability estimates. For these scores, with items of low discrimination, variability decreased as the number of items increased. Better initial ability estimates were associated with decreased variability for the first two scores, but for the Bayesian ability score variability increased with increasing numbers of items.

Variability of score 6, the Bayesian error predictor score, increased with increasing item discriminations. Scores 4 and 5, the average difficulty error predictor scores, showed only slight decreases with improved discriminations. All error predictor scores showed decreasing variabilities as test length increased. Scores 4 and 5 showed decreasing variability as the initial ability estimates improved, but score 6 showed a slight increase.

Table 9
Standard Deviations of Scores for Fixed-Length Stradaptive as a Joint Function of Item Discriminations (a) and Validity of Initial Ability Estimates

a	No. Items	Fixed Entry	Initial Ability Correlation 0.0	0.5	1.0	a	No. Items	Fixed Entry	Initial Ability Correlation 0.0	0.5	1.0
			Score 1						Score 4		
0.5	10	.844	.949	1.020	1.090	0.5	10	.101	.144	.115	.093
	20	.878	.914	.995	1.050		20	.067	.078	.068	.069
	40	.882	.891	.947	.988		40	.042	.046	.045	.044
	60	.894	.882	.899	.951		60	.032	.033	.034	.034
1.0	10	.838	.882	.984	1.060	1.0	10	.095	.134	.102	.084
	20	.893	.911	.967	1.040		20	.056	.071	.060	.052
	40	.913	.908	.967	.991		40	.035	.041	.035	.033
	60	.927	.938	.951	.977		60	.023	.026	.025	.024
2.0	10	.871	.897	.996	1.040	2.0	10	.095	1.530	.096	.072
	20	.912	.935	.990	1.010		20	.056	.074	.057	.053
	40	.944	.942	.979	1.020		40	.035	.040	.035	.033
	60	.952	.937	.974	.983		60	.039	.029	.024	.022
			Score 2						Score 5		
0.5	10	.790	.895	.976	1.060	0.5	10	.060	.076	.066	.055
	20	.841	.873	.961	1.020		20	.041	.048	.044	.045
	40	.846	.849	.908	.948		40	.026	.029	.030	.030
	60	.856	.844	.866	.911		60	.020	.021	.022	.023
1.0	10	.793	.830	.959	1.050	1.0	10	.051	.069	.056	.049
	20	.869	.876	.941	1.020		20	.032	.042	.036	.032
	40	.889	.883	.945	.968		40	.020	.024	.020	.021
	60	.901	.913	.926	.957		60	.014	.016	.015	.015
2.0	10	.824	.851	.966	1.030	2.0	10	.051	.074	.056	.042
	20	.886	.899	.964	1.000		20	.029	.042	.031	.029
	40	.921	.919	.957	.998		40	.018	.023	.020	.019
	60	.931	.918	.953	.961		60	.032	.016	.014	.014
			Score 3						Score 6		
0.5	10	.687	.682	.689	.665	0.5	10	.044	.048	.054	.059
	20	.791	.796	.795	.792		20	.040	.042	.046	.050
	40	.885	.884	.896	.899		40	.029	.030	.032	.033
	60	.929	.913	.882	.922		60	.024	.024	.024	.026
1.0	10	.833	.824	.864	.855	1.0	10	.063	.079	.090	.106
	20	.908	.917	.923	.937		20	.047	.053	.057	.061
	40	.943	.946	.972	.969		40	.029	.031	.033	.034
	60	.971	.980	.968	.977		60	.021	.022	.023	.023
2.0	10	.934	.928	.952	.912	2.0	10	.078	.101	.120	.149
	20	.949	.973	.966	.960		20	.050	.051	.059	.060
	40	.975	.982	.982	1.000		40	.024	.024	.025	.025
	60	.985	.970	.995	.985		60	.017	.017	.016	.016

Correlational Analyses

Table 10 shows a matrix of intercorrelations among Fixed-Length Stradaptive test scores. As the table shows, the three ability scores correlated very highly; the two average difficulty scores, scores 1 and 2, correlated almost perfectly (i.e., .999) and the Bayesian score correlated .992 and .993 with them. Scores 4 and 5, the error predictor scores corresponding to the

Table 10
Intercorrelations Among Fixed-Length Stradaptive Scores

	Ability Scores			Error Predictor Scores		
	1	2	3	4	5	6
1	1.000	.999	.992	-.079	-.169	.450
2	.999	1.000	.993	-.090	-.173	.446
3	.992	.993	1.000	-.166	-.246	.407
4	-.079	-.090	-.166	1.000	.958	.351
5	-.169	-.173	-.246	.958	1.000	.327
6	.450	.446	.407	.351	.327	1.000

average difficulty scores, correlated highly with each other (.958) but low with the Bayesian error predictor score (.351 and .327, respectively). Error predictor scores 4 and 5 correlated very slightly and negatively with the three ability scores, but score 6, the Bayesian error predictor score, correlated moderately with all three ability scores.

The error predictor scores were designed to be independent of the ability scores. The fact that the Bayesian error score correlated with the ability scores suggests that it might not be as useful as the others as a measure of error. This result may be due to the guessing probability being non-zero in the model, which may have affected the Bayesian error score.

Table 11 shows the correlations of the three Fixed-Length Stradaptive ability scores with the generating ability. These correlations increased for all scores as the item discriminating power and test length increased. For the average difficulty scores, scores 1 and 2, these correlations increased as the validity of the initial ability estimate increased. As can be seen, this trend diminished as test length increased. No definite trend with respect to initial ability estimates was observed in the Bayesian score. This is because a constant prior (i.e., population parameters) was given to the Bayesian scoring routine while the average difficulty scores implicitly incorporated the initial ability estimate information. The capability to explicitly use prior information could easily be added to the Bayesian score but this capability is not without its disadvantages (e.g., the goodness of the prior must be stated explicitly and this will allow a poor prior to bias the score).

Table 11
Score-Ability Correlations for Fixed-Length Stradaptive as a Joint Function of Item Discrimination (a), Test Length, and Validity of Initial Ability Estimates

a	No. Items	Fixed Entry	Initial Ability Correlation 0.0	0.5	1.0
		Score 1			
0.5	10	.656	.568	.719	.831
	20	.767	.721	.797	.857
	40	.856	.846	.866	.889
	60	.906	.901	.903	.921
1.0	10	.813	.719	.847	.901
	20	.905	.873	.905	.924
	40	.940	.937	.952	.958
	60	.965	.962	.966	.968
2.0	10	.885	.809	.898	.939
	20	.946	.920	.949	.961
	40	.970	.966	.973	.977
	60	.980	.978	.982	.984
		Score 2			
0.5	10	.649	.552	.720	.836
	20	.765	.720	.798	.857
	40	.853	.844	.863	.886
	60	.906	.899	.901	.919
1.0	10	.824	.721	.844	.905
	20	.908	.877	.906	.925
	40	.943	.939	.953	.958
	60	.965	.961	.966	.968
2.0	10	.896	.809	.903	.945
	20	.949	.923	.952	.965
	40	.972	.966	.974	.978
	60	.980	.979	.983	.985
		Score 3			
0.5	10	.689	.683	.667	.678
	20	.798	.782	.794	.794
	40	.869	.872	.869	.871
	60	.920	.912	.907	.917
1.0	10	.840	.840	.844	.838
	20	.918	.919	.915	.909
	40	.955	.954	.959	.957
	60	.971	.970	.972	.972
2.0	10	.919	.905	.916	.918
	20	.963	.958	.963	.967
	40	.983	.984	.985	.985
	60	.989	.989	.990	.990

Correlations of scores 1 and 2 with generating ability were higher when using a fixed entry point than when using a variable entry point with invalid prior information. A similar result was observed, in general, for the Bayesian score. Score-ability correlations for scores 1 and 2 using a variable entry point were, in general, higher than those using a fixed entry point when the initial ability estimate correlated .5 with generating ability. When initial ability estimates correlated 1.0 with actual ability, score-ability correlations for scores 1 and 2 were always higher than with a fixed entry point. But, in general, the advantage of prior information diminished as test length increased. Score 3, the Bayesian score, usually resulted in higher score-ability correlations when a fixed entry point was used than when prior information was available regarding a testee's ability level.

Of the two average difficulty scores, score 2, the average difficulty of all items administered, correlated higher with generating ability than did score 1 for tests with items having discriminations of 1.0 or 2.0. Score 1, the average difficulty of all items answered correctly, correlated higher with ability for tests with items having discriminations of .5. Score 3, the Bayesian score, correlated higher with ability than either of the average difficulty scores when no prior information was available, regardless of item discriminations. In general, score 3 also correlated as high or higher than did the average difficulty scores for tests having 40 or 60 items or item discriminations of 2.0 even when prior information was available.

Comparison with the conventional test. The fairest general comparison between Fixed-Length Stradaptive scores and conventional test scores is with no prior information, since the conventional test cannot use prior information. The best Fixed-Length Stradaptive score for those conditions was the Bayesian score.

Comparing Tables 1 and 11, when discriminations were .5, the conventional test correlated higher with ability for all lengths shorter than 60 items. When discriminations were 1.0, the stradaptive test correlated higher at all lengths greater than 10 items. When discriminations were 2.0, the stradaptive test correlated higher with ability than did the conventional test at all of the four lengths investigated. For the 10-item tests, the stradaptive score 3 correlation with ability was .919, while that for the conventional test was .888; at 60 items the respective correlations were .989 and .926.

Although the comparison with the fixed entry point is of most interest in a pure research situation, it is also appropriate to compare the usual modes of implementing the two testing strategies, i.e., the conventional test without prior information and the stradaptive test with a moderately valid prior ability estimate (i.e., .50). Under these circumstances, score-ability correlations were higher for the stradaptive test using all scoring methods with highly discriminating items. Using moderately discriminating items, stradaptive correlations were higher, in general, for tests longer than ten items. For items with low (a=.5) discriminations, the conventional test scores correlated higher with ability.

Information Analysis

Table 12 presents average heights of information curves for the Fixed-Length

Table 12
Average Information Provided by Revised Stradaptive Scores as a Joint Function of Item Discriminations (a), Test Length, and Validity of Initial Ability Estimates

a	No. Items	Fixed Entry	Initial Ability Correlation 0.0	0.5	1.0
			Score 1		
0.5	10	.645	.343	.924	1.943
	20	1.412	.913	1.642	2.442
	40	2.823	2.260	2.989	3.708
	60	4.234	3.721	4.595	5.191
1.0	10	1.864	.823	2.034	3.990
	20	3.762	2.247	4.122	5.825
	40	8.221	6.049	8.494	10.194
	60	12.835	9.963	12.721	15.262
2.0	10	4.460	1.413	3.530	8.184
	20	9.360	4.510	7.833	13.286
	40	20.587	11.705	17.836	23.958
	60	29.942	18.725	26.402	32.924
			Score 2		
0.5	10	.601	.312	.919	2.077
	20	1.368	.902	1.647	2.485
	40	2.779	2.299	2.999	3.701
	60	4.220	3.809	4.625	5.170
1.0	10	1.805	.763	2.039	4.349
	20	3.861	2.319	4.281	6.211
	40	8.608	6.350	8.937	10.808
	60	13.468	10.721	13.611	16.432
2.0	10	4.447	1.311	3.540	9.459
	20	10.180	4.557	8.282	15.414
	40	23.014	12.729	19.574	28.061
	60	33.283	20.727	29.254	37.862
			Score 3		
0.5	10	.789	.748	.833	.847
	20	1.721	1.586	1.719	1.797
	40	3.345	3.328	3.348	3.474
	60	4.989	5.027	5.141	5.279
1.0	10	2.204	1.951	2.444	2.496
	20	4.565	4.579	5.155	5.268
	40	10.432	10.065	10.994	11.054
	60	16.643	16.357	16.948	17.059
2.0	10	5.100	4.091	4.885	5.702
	20	12.354	11.887	12.271	13.139
	40	29.151	28.698	30.753	33.455
	60	46.748	44.496	48.183	50.570

Stradaptive scores. As was observed with the score-ability correlations, average information increased with item discrimination and test length for all scoring methods.

Whereas score-ability correlations increased with higher validity of the initial ability estimate only for scores 1 and 2, average information showed that increasing trend for all three scores. An advantage of the Bayesian score with respect to this trend is observed in Table 12. Since, as implemented, the Bayesian scoring procedure used a constant prior regardless of the testee's entry point, the effects of poor prior information resulting in an inappropriate entry point were negligible, whereas the effect on an average difficulty score was substantial. As an example, the average information from a 60-item stradaptive test provided by score 2 dropped from 33.283 to 20.727 as a fixed entry point changed to a random entry point which was uncorrelated with actual ability. Under the same conditions, the average information provided by the Bayesian score dropped only slightly from 46.748 to 44.496.

Score 2 generally provided higher levels of information than did score 1 when item discriminations were 1.0 or 2.0. Although score 1 correlated higher with ability than did score 2 when item discriminations were low (a=.5), this result did not occur when average information values were compared. As with score-ability correlations, the Bayesian score provided the highest level of average information of the three scores when no prior information was available. Furthermore, it provided the highest level of information when initially ability estimates correlated .5 with ability, except for a 10-item test with item discriminations of .5. Even when the average difficulty scores had prior information correlating 1.0 with ability and the Bayesian score used no prior information, the latter scoring method provided a higher level of average information when test length was 60 and the items were moderately or highly discriminating.

Table 13 presents the coefficients of variation for the height of the information curves of the three Fixed-Length Stradaptive ability scores. One trend was apparent in all three scores: equiprecision decreased (and hence the coefficient of variation increased) as the items became more discriminating. For example, under the fixed entry point condition, score 1 coefficients increased from an average of about 34 to an average of about 64 as discriminations increased from .5 to 2.0.

No consistent trends were apparent with respect to test length, although equiprecision appeared to improve for all scores as test length increased when prior information was either not used (i.e., fixed point entry) or was very poor (i.e., correlation 0.0 with ability). Coefficients of variation for scores 1 and 2 showed a U-shaped quadratic trend with respect to improving initial ability estimates. This trend is not easily explained, did not generally appear using the Bayesian score (score 3), and may involve a more complex interaction with test length. For example, the trend was readily apparent for score 1 with a length of ten items and discriminations of .5, but flattened out somewhat, under the same conditions, for score 2. Coefficients of variation for the Bayesian score, instead of showing this trend, showed a monotonic decreasing trend with increasing goodness of initial ability estimates within levels of item discrimination and test length.

Table 13
Coefficients of Variation for Information Functions of Fixed-Length Stradaptive Scores as a Joint Function of Item Discriminations (a), Test Length, and Validity of Initial Ability Estimates

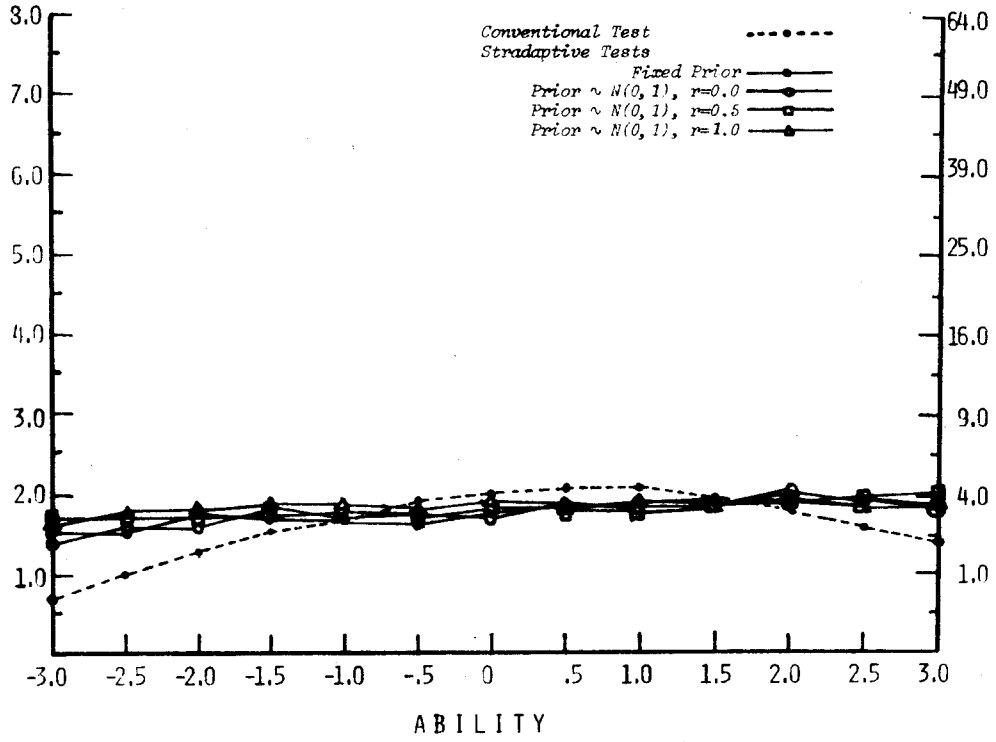
a	No. Items	Fixed Entry	Initial Ability Correlation 0.0	0.5	1.0
			Score 1		
0.5	10	40.218	36.915	25.681	42.712
	20	36.415	32.185	21.876	33.565
	40	31.192	27.389	24.808	29.355
	60	29.637	27.206	27.595	29.680
1.0	10	53.558	60.132	36.237	32.839
	20	50.029	49.366	34.918	37.824
	40	46.274	41.329	35.869	36.683
	60	44.351	39.448	38.758	37.147
2.0	10	65.532	69.094	47.545	45.275
	20	62.176	57.474	47.305	56.363
	40	62.566	49.879	46.948	55.433
	60	65.565	47.057	51.949	60.603
			Score 2		
0.5	10	32.132	36.367	17.789	42.311
	20	26.793	31.372	16.064	31.008
	40	22.350	24.290	16.027	25.036
	60	20.002	22.160	18.417	23.002
1.0	10	49.193	65.831	28.147	32.209
	20	44.836	53.478	26.024	34.334
	40	37.983	40.565	26.867	30.773
	60	35.312	37.427	30.882	29.794
2.0	10	64.505	82.484	45.660	41.149
	20	59.592	66.010	43.873	52.664
	40	60.093	55.599	41.947	51.329
	60	60.594	51.547	46.137	54.752
			Score 3		
0.5	10	26.370	26.023	15.656	12.109
	20	22.003	20.847	11.664	8.255
	40	16.969	17.288	12.651	11.397
	60	14.848	15.880	14.130	13.401
1.0	10	40.828	42.653	18.994	14.132
	20	33.230	36.038	17.937	10.805
	40	24.806	25.312	17.257	13.561
	60	21.786	22.618	15.961	11.990
2.0	10	54.820	56.473	28.118	26.204
	20	42.800	45.162	27.990	14.572
	40	38.263	38.061	30.573	17.805
	60	34.114	34.516	27.272	24.948

Comparison with the conventional test. Comparing information provided by a fixed entry point stradaptive test using the Bayesian score with the information provided by a conventional test (Table 1), the stradaptive test always provided a higher average level of information. At no combination of test length and item discrimination did the information function of the conventional test have a higher average value than that of score 3 of the stradaptive test. Stradaptive scores 1 and 2 (with fixed entry) also provided higher average levels of information than did the conventional test, for all test with item discriminations of 1.0 or 2.0. When the stradaptive test utilized valid prior information, its average level of information exceeded that of the conventional test across all test lengths, levels of discrimination, and scoring methods.

Equiprecision of measurement provided by the Fixed-Length Stradaptive test was superior to the equiprecision provided by the conventional test in the vast majority of comparisons. Equiprecision provided by the Fixed-Length Stradaptive Bayesian score was superior to that of the conventional test score in all cases where test lengths and item discriminations were matched. The lowest coefficient of variation (i.e., the best equiprecision) generated by the conventional test was 39.617, for a 60-item test with items of .5 discrimination. The Bayesian score of Fixed-Length Stradaptive provided better equiprecision than this in all but six out of 48 conditions; in each case when the conventional test was more equiprecise, the stradaptive test was composed of 20 items or fewer.

Figure 1

Information Curves for 40-Item Conventional and Stradaptive Tests Using Items with Discriminations of a=0.5



In general, the use of valid prior information within Fixed-Length Stradaptive resulted in greater equiprecision of measurement. For score 1, the most equi-

precise measurement was observed for initial ability correlations of .5. For scores 2 and 3, increasingly valid entry point information resulted in more equiprecise measurement with items of moderate (a=1.0) and high discriminations (a=2.0), and for low discriminating items (a=.5) for score 3. With a few exceptions, a fixed entry point was better than an invalid (r=0.0) variable entry point.

Graphic comparison of information curves. Figures 1, 2 and 3 show graphically the effect of different item parameters and different validities of prior information on the information curves of the Stradaptive Bayesian score; for comparison purposes, these figures also include information curves for conventional test scores. Figure 1 shows information curves based on item discriminations of a=.5 for a 40-item conventional test and for 40-item stradaptive tests with different initial ability estimate validities. With such low item discriminations, all tests resulted in very low and flat information curves.

Figure 2

Information Curves for 40-Item Conventional and Stradaptive Tests Using Items with Discriminations of a=1.0

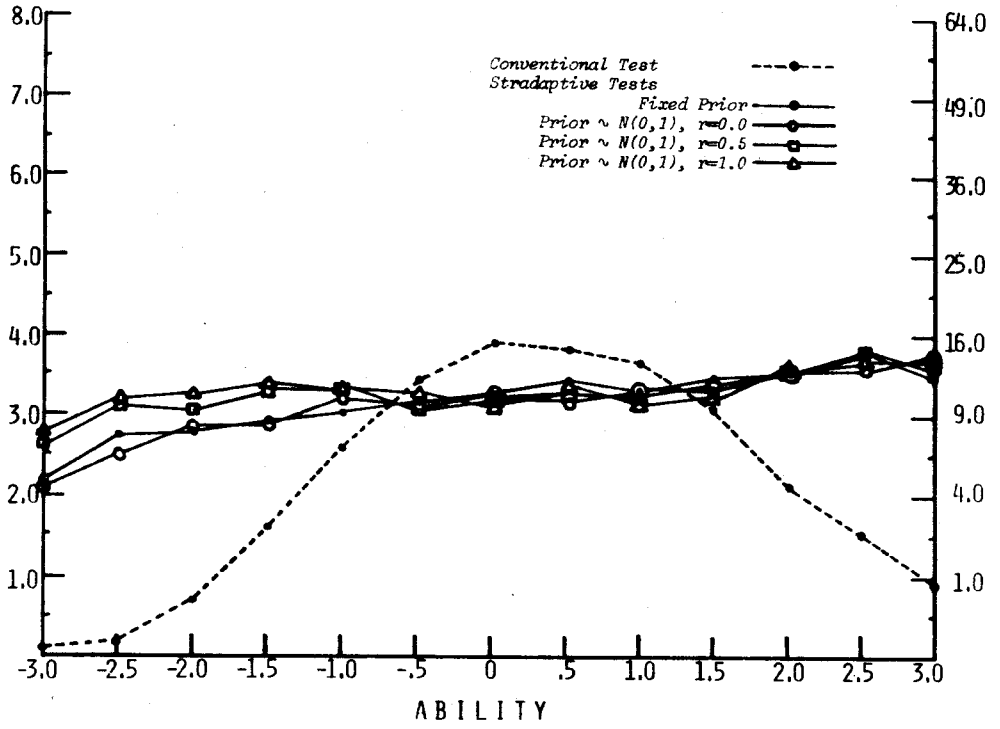
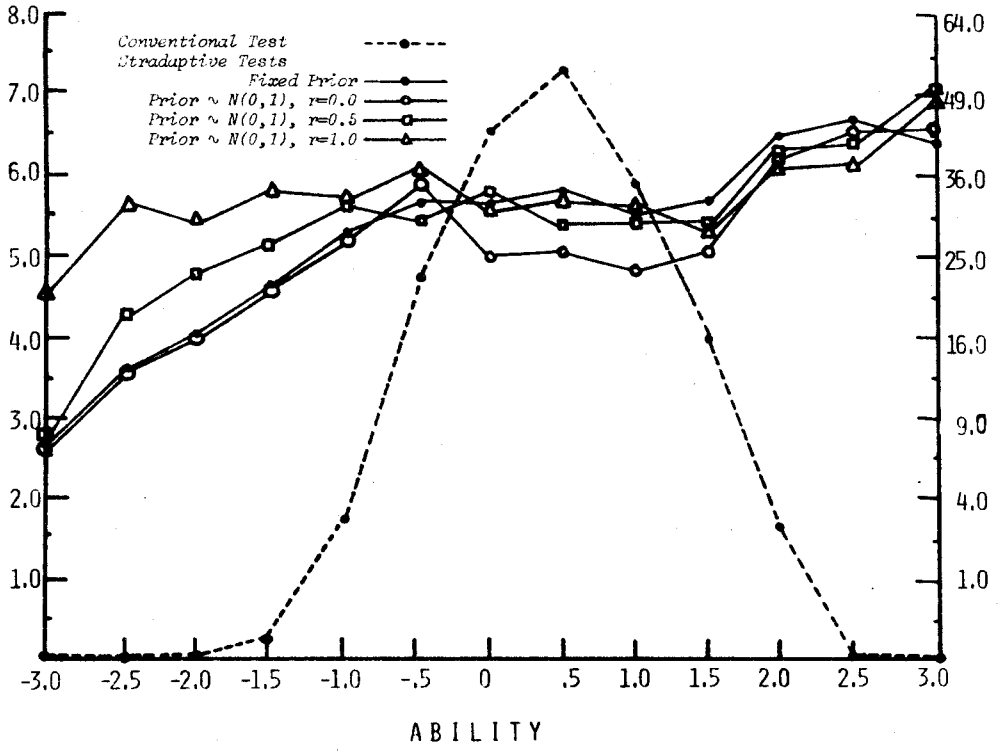


Figure 2 shows the same curves for tests with items having discriminations of a=1.0. Several trends are apparent: 1) the conventional test provided better measurement than the stradaptive tests in the middle range of the ability range, but less precise measurement in the extremes; 2) the stradaptive tests provided more equiprecise measurement (i.e., flatter curves) than did the conventional test; and 3) the initial ability estimates had an effect primarily at low ability levels.

Figure 3 shows the curves for tests using very discriminating items (i.e., a=2.0). The information curves were again higher and the three observations made for Figure 2 were even more obvious.

Figure 3

Information Curves for 40-Item Conventional and Stradaptive Tests Using Items with Discriminations of a=2.0



The general upward trend in the curves observed in Figures 2 and 3 results from the .20 guessing parameter used in the simulations. It is interesting to note, however, that the maximum values of information achieved by the stradaptive test at the high ability levels are essentially equal to the highest information values achieved by the conventional test.

Termination Criterion Analysis

Table 14 shows correlations of the three error predictor scores with absolute deviations from the line of relations between standardized ability and ability scores. Error scores and ability scores in Table 14 are ordered such that the error scores correspond to the ability score on the principal diagonal of the correlation matrix.

The data in Table 14 show that: 1) the correlations increased as the item discriminating powers increased; 2) the error scores were not necessarily most predictive of the corresponding ability score on the diagonal; and 3) the correlations were, in general, quite low. When item discriminations were .5, the

correlations were so low as to suggest that no information about precision of measurement is provided by the error scores. The correlations of .205 and .250 suggest that when item discriminations are 1.0, error scores 4 and 5 might be of slight utility in predicting precision of measurement. The Bayesian error score, score 6, was slightly correlated with errors of the two average difficulty ability scores but essentially unpredictive for errors of the Bayesian ability score, score 3. When discriminations were 2.0, error scores 4 and 5 were slightly more predictive of errors in their corresponding ability scores than they were when discriminations were 1.0. The Bayesian error score was still the least predictive of the three.

Table 14
Correlations Between Error Predictor Scores and Absolute Deviations from the Line of Relations Between Ability and Ability Scores, as a Function of Item Discriminations (a)

a	Ability Score	Error Scores 4	5	6
0.5	1	-.020	-.013	-.052
	2	-.018	-.005	-.045
	3	-.028	-.038	-.004
1.0	1	.205	.231	.168
	2	.221	.250	.187
	3	.013	.007	.062
2.0	1	.350	.398	.191
	2	.323	.364	.149
	3	.247	.249	.179

Using the correlation of error scores with ability score deviations as an evaluative criterion, it appears that none of the three criteria investigated are very useful for estimating the magnitude of measurement errors. However, a different criterion might provide different results. Furthermore, no trait of inconsistency or unpredictability was programmed into the response model. The analysis, therefore, was sensitive only to imprecision introduced through fortuitous response instability affecting the psychometric properties of the tests. Live testees might be predictably inconsistent and thus demonstrate better validities for the error scores. A future study might compare shapes of information curves provided by tests terminated under these criteria as well as the original criterion suggested by Weiss (1973). The best termination criterion would, in that situation, be the one which produced the most equiprecise measurement at all levels of ability. Such a study might also investigate models of a trait of inconsistency or unpredictability.

SUMMARY AND CONCLUSIONS

Conventional Test

Two psychometric characteristics of the conventional test are worthy of note. First, the score-ability correlation is not a monotonically increasing

function of item discriminating power. For conventional tests longer than 20 items, the score-ability correlation decreased as items became more discriminating than a=1.0. The second observation is that equiprecision of measurement with a conventional test is a monotonic decreasing function of item discrimination. These two observations combined suggest that if sufficiently discriminating items are available, the conventional test provides precise measurement for so few people that improving the quality of the items does not improve the quality of the measurement.

Variable-Length Stradaptive

The original form of the stradaptive test showed characteristics in this study similar to its characteristics as derived from previous live-testing studies. Means and standard deviations of scores obtained using the real item pool were slightly different in the simulation data. This was probably due to the use of item parameters based on small subject groups in the live-testing study. The difference in results may, however, have been a function of the failure of aspects of the simulation model to adequately reflect the behavior of real testees.

The length of the Variable-Length Stradaptive test shortened substantially as items became more discriminating, but showed no trend with improving initial ability estimates. Tests using the real item pool were shorter than would have been expected considering the discriminations of the item pool. This supports the suggestion that, under Weiss' original ceiling stratum termination criterion, test length may be decreased considerably by putting the most discriminating items at the beginnings of each of the strata.

The same clusters of scores observed in live-testing studies were observed in the simulation data; and, as in the live-testing data, the average difficulty scores, scores 8 and 9, had the highest indices of reliability (i.e., correlations with generating ability). No definite trend in score-ability correlations was observed as a function of the quality of the initial ability estimates. This suggests that variable entry to Variable-Length Stradaptive does not increase its capability of reflecting true ability level, on the average. An increasing trend in score-ability correlations with increasing item discriminations was noted. This suggests that there is a deficiency in the termination criterion since it does not keep precision of measurement constant, as it was intended to do.

Variable-Length Stradaptive vs. Conventional

Comparing Variable-Length Stradaptive to the conventional test, the best stradaptive score, score 8, had higher score-ability correlations only when item discriminations were higher than a=1.0. The same observation was true for average information. Better equiprecision of measurement, on the other hand, was always provided by the stradaptive test, with coefficients of variation for information functions of the conventional test sometimes being more than twice as large as those of the stradaptive test operating under the same conditions.

The simple question of which of the two testing strategies is better cannot be answered without specifying criteria and conditions. The Variable-Length Stradaptive testing strategy always provided more equiprecise measurement than the conventional test, but provided more average information and higher score-ability correlations only when the items were highly discriminating.

Fixed-Length Stradaptive

Intercorrelations among Fixed-Length Stradaptive scores revealed that scores 1 and 2, respectively the average difficulty of items correct and administered, were nearly identical, correlating .999. The Bayesian ability score also correlated highly with the first two (r=.993 and .992). Error predictor scores 4 and 5, the error scores corresponding to ability scores 1 and 2, correlated highly among themselves, moderately with the Bayesian error score, and poorly with the ability scores. The Bayesian error score correlated higher with the ability scores than with the other error scores.

Ability scores 1 and 2, for all practical purposes, performed equally well in terms of score-ability correlations, average information, and equiprecision. The Bayesian ability score performed better than the first two scores, in terms of score-ability correlations and average information, when tests were more than 20 items long or when initial ability correlations were less than .5. It did not perform as well in other conditions because it was not explicitly given prior information when prior information was available and could not use it implicitly as could the average difficulty scores. Given an informative prior the Bayesian score would probably always be superior to the average difficulty scores. An advantage of the Bayesian score when given only a population prior is that the effects of poor prior information resulting in an inappropriate entry point are negligible whereas the effect on the average difficulty scores is great. Although no score was best with respect to average information under the conditions studied, the Bayesian ability score always provided more equiprecise measurement than did the average difficulty scores.

Fixed-Length Stradaptive vs. Conventional

When compared to the conventional test, the Bayesian score, which was generally the best score of Fixed-Length Stradaptive, correlated higher with ability than did the conventional test score when tests were long or discriminations were high. For item discriminations of a=.5, the stradaptive test's Bayesian score correlated higher with ability than did the conventional test score when the test length was 60 items. When item discriminations were 1.0, the Bayesian score correlated higher when tests were longer than 10 items. When discriminations were 2.0, the Bayesian score correlated higher with ability than did the conventional test at all test lengths. The Bayesian score of Fixed-Length Stradaptive always provided higher average information and better equiprecision than did the conventional test when item discriminations and test lengths were equated.

Termination Criteria

Error predictor scores investigated in conjunction with Fixed-Length Stradaptive appeared to provide little information about errors of measurement, although slight correlations with absolute error from the line of relations were observed when item discriminations were high. These data do not lend support to the idea of using these error scores as termination criteria. However, an alternative approach for future research on termination criteria might be to terminate tests on the basis of the various criteria and compare shapes of the resulting information curves. The best termination criterion would produce the flattest information curve.

Conclusions

The data support the contention that the stradaptive testing strategy can produce better measurement than comparable conventional tests in terms of amount of information provided, equality of information provided at different ability levels, and in some conditions, in terms of correlations of scores with ability. These advantages become even greater as item discriminations improve. The data further suggest that 1) the Bayesian scoring technique is a very good method for scoring the stradaptive test; 2) the use of prior information to provide variable entry points into the fixed-length stradaptive test generally improves the measurement characteristics of the resulting scores; and 3) further research is needed to develop and refine flexible termination criteria for the stradaptive testing strategy.

APPENDIX A

A Fortran IV Bayesian Scoring Routine

```
      SUBROUTINE BSCOR (BTHET,BVAR,DIF,DIS,IRESP)
C---- CALLING PARAMETERS
C----      BTHET :  MEAN OF PRIOR ABILITY DISTRIBUTION
C----      BVAR  :  VARIANCE OF PRIOR ABILITY DISTRIBUTION
C----      DIF   :  B-VALUE OF ITEM
C----      DIS   :  A-VALUE OF ITEM
C----      IRESP :  RESPONSE -- 1 = CORRECT, 0 = INCORRECT
C---- GUESSING PARAMETER SET TO 0.2 HERE
      GUESP=0.2
      D=(DIF-BTHET)/SQRT(2.0*(1.0/DIS**2+BVAR))
      ERFD=2.0*CDFN(D*1.41421)-1.0
      EDSQ=EXP(D**2)
      EDSQI=1.0/EDSQ
      XKINV=0.5*(1.0-ERFD)
      XLINV=GUESP+(1.0-GUESP)*XKINV
      XL=1.0/XLINV
      IF (IRESP .NE. 1) GO TO 10
      S=0.398942*(SQRT(BVAR)/SQRT(1.0+(1.0/DIS**2)*
     +1.0/BVAR))*(1.0/XKINV)*EDSQI
      T=1.0-1.772454*D*EDSQ*(1.0-ERFD)
      BTHET=BTHET+(1.0-GUESP)*XKINV*XL*S
      BVAR=BVAR-(1.0-GUESP)*XKINV*XL*S**2*(T-GUESP*XL)
      RETURN
   10 BTHET=BTHET-0.797885*(BVAR/SQRT(1.0/DIS**2+
     +BVAR))*EDSQI*(1.0/(1.0+ERFD))
      PART1=1.128379/(1.0+(1.0/DIS**2)*(1.0/BVAR))
      PART2=1.0/(EDSQ*(1.0+ERFD))**2
      PART3=0.564190+D*EDSQ*(1.0+ERFD)
      BVAR=BVAR*(1.0-PART1*PART2*PART3)
      RETURN
      END
```

APPENDIX B

Table B-1

Item Parameters for the Real Stradaptive Item Pool

	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
Mean	-2.648	1.290	-1.951	.830	-1.314	.749	-.666	.617	-.020	.643	.695	.528	1.313	.460	2.006	.482	2.621	.427
S.D.	.176	.925	.212	.398	.202	.263	.178	.236	.199	.220	.206	.175	.182	.115	.206	.133	.273	.163
High	-2.393	3.000	-1.657	1.756	-1.013	1.396	-.343	1.822	.329	1.306	.977	.980	1.630	.718	2.313	.851	3.113	.840
Low	-2.980	.406	-2.322	.313	-1.653	.317	-.998	.301	-.319	.317	.337	.310	1.004	.312	1.649	.339	2.320	.214
	-2.415	3.000	-1.989	1.756	-1.509	1.396	-.703	1.822	-.054	1.306	.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	1.427	-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
	-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	.511	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	1.007	-1.657	.927	-1.653	.822	-.813	.665	.172	.772	.784	.448	1.004	.418	2.265	.352		
	-2.865	1.007	-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	-1.884	.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	-1.485	.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				
	-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	.000	.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	.582	.329	.527								
	-2.578	.570	-1.804	.418	-1.074	.555	-.725	.568	-.233	.505								
	-2.980	.559	-1.825	.417	-1.020	.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	.511	-.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	.338	-.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										

Table B-2
Difficulties (b), by Stratum, of the Hypothetical Stradaptive Item Pool Having Item Discriminations of a =.5

Stratum								
1	2	3	4	5	6	7	8	9
-2.504	-1.781	-1.448	-.559	-.119	.663	1.337	1.898	2.967
-2.695	-1.792	-1.373	-.683	.285	.838	1.252	1.789	2.366
-2.863	-2.271	-1.152	-.835	.249	.617	1.354	2.102	2.941
-2.497	-1.929	-1.271	-.898	-.112	.513	1.488	1.864	2.379
-2.338	-1.907	-1.470	-.508	.221	.451	1.582	1.693	2.637
-2.377	-1.756	-1.411	-.386	-.123	.444	1.647	1.788	2.575
-2.520	-2.115	-1.329	-.333	-.288	.584	1.115	1.929	2.642
-2.588	-2.029	-1.381	-.569	.171	.380	1.481	2.146	2.870
-2.989	-2.306	-1.637	-.393	-.038	.764	1.024	2.054	2.872
-2.709	-2.160	-1.428	-.863	.026	.740	1.403	2.185	2.511
-2.658	-1.895	-1.321	-.773	.227	.739	1.061	1.986	2.573
-2.993	-2.234	-1.140	-.680	.020	.603	1.165	1.821	2.754
-2.447	-2.307	-1.288	-.886	-.234	.929	1.640	2.310	2.332
-2.350	-1.664	-1.425	-.774	.072	.387	1.179	2.310	2.845
-2.607	-2.283	-1.127	-.816	-.186	.412	1.193	2.273	2.735
-2.372	-1.873	-1.209	-.560	.097	.512	1.199	2.229	2.701
-2.721	-1.909	-1.084	-.646	.291	.948	1.453	1.791	2.358
-2.640	-2.277	-1.401	-.791	.014	.726	1.636	2.163	2.908
-2.369	-1.938	-1.407	-.460	-.207	.944	1.511	2.120	2.457
-2.735	-2.287	-1.588	-.880	-.259	.777	1.583	2.029	2.711
-2.472	-1.784	-1.024	-.343	.218	.820	1.121	1.837	2.948
-2.909	-1.989	-1.357	-.912	-.146	.415	1.454	1.744	2.698
-2.611	-2.201	-1.502	-.377	.164	.995	1.355	2.198	2.409
-2.742	-2.136	-1.615	-.970	.271	.728	1.659	1.893	2.784
-2.369	-1.932	-1.109	-.647	.207	.546	1.305	1.779	2.765
-2.487	-2.142	-1.291	-.467	.250	.795	1.292	1.722	2.885
-2.666	-1.730	-1.411	-.566	.266	.564	1.384	1.836	2.591
-2.393	-2.090	-1.292	-.829	-.292	.365	1.425	1.819	2.919
-2.413	-1.681	-1.337	-.889	.258	.796	1.344	2.286	2.476
-2.714	-2.064	-1.206	-.798	-.117	.526	1.151	2.098	2.801
-2.531	-1.867	-1.135	-.358	.100	.991	1.598	1.705	2.334
-2.359	-2.148	-1.279	-.458	.033	.358	1.542	1.807	2.447
-2.603	-1.703	-1.051	-.397	-.289	.898	1.239	2.019	2.753
-2.920	-1.756	-1.099	-.563	.074	.948	1.030	2.202	2.444
-2.879	-1.941	-1.333	-.421	.327	.937	1.338	1.804	2.978
-2.638	-2.311	-1.575	-.527	-.076	.706	1.349	1.859	2.708
-2.940	-1.996	-1.568	-.513	-.114	.854	1.649	2.316	2.925
-2.435	-1.887	-1.632	-.433	.161	.721	1.590	2.034	2.486
-2.368	-1.952	-1.345	-.757	.285	.401	1.264	2.133	2.511
-2.561	-1.909	-1.520	-.365	.042	.741	1.507	1.974	2.362
-2.754	-1.892	-1.570	-.566	.122	.664	1.256	1.839	2.992
-2.599	-1.788	-1.485	-.407	.144	.515	1.511	2.264	2.925
-2.338	-1.901	-1.171	-.695	-.025	.723	1.376	2.022	2.945
-2.799	-2.281	-1.115	-.643	-.274	.462	1.355	2.151	2.749
-2.714	-2.061	-1.056	-.556	-.240	.997	1.377	1.928	2.955
-2.742	-2.129	-1.317	-.486	.025	1.000	1.452	2.221	2.422
Mean -2.607	-2.000	-1.332	-.614	.032	.683	1.374	1.999	2.681
S.D. .190	.192	.171	.186	.190	.200	.176	.190	.213

Table B-3

Difficulties (b), by Stratum, of the Hypothetical Stradaptive Item Pool Having Item Discriminations of a =1.0

	Stratum								
	1	2	3	4	5	6	7	8	9
	-2.653	-1.696	-1.616	-.742	-.093	.672	1.095	2.294	2.344
	-2.707	-2.032	-1.081	-.847	.272	.558	1.023	1.729	2.837
	-2.395	-2.081	-1.424	-.412	-.111	.824	1.657	1.790	2.959
	-2.483	-1.896	-1.418	-.807	-.305	.832	1.046	1.798	2.764
	-2.478	-1.948	-1.237	-.977	.216	.865	1.450	2.327	2.721
	-2.642	-1.787	-1.141	-.783	.246	.787	1.414	1.861	2.382
	-2.650	-2.050	-1.468	-.343	.146	.713	1.053	2.148	2.384
	-2.931	-1.787	-1.323	-.691	-.136	.339	1.386	2.171	2.445
	-2.659	-1.802	-1.210	-.437	.256	.954	1.042	2.172	2.701
	-2.430	-1.981	-1.056	-.448	-.214	.745	1.537	2.308	2.367
	-2.925	-1.957	-1.534	-.504	-.195	.833	1.336	1.982	2.416
	-2.898	-1.782	-1.414	-.638	-.070	.580	1.127	1.854	2.901
	-2.526	-1.928	-1.533	-.462	.029	.937	1.191	1.829	2.953
	-2.743	-1.945	-1.488	-.510	.214	.430	1.392	1.729	2.792
	-2.552	-1.800	-1.619	-.896	.186	.513	1.185	1.963	2.385
	-2.383	-1.882	-1.030	-.648	-.106	.882	1.591	2.186	2.792
	-2.750	-2.327	-1.546	-.383	.193	.816	1.237	1.967	2.807
	-2.908	-1.679	-1.114	-.970	.208	.513	1.639	2.134	2.986
	-2.869	-2.251	-1.066	-.849	-.031	.423	1.262	2.091	2.962
	-2.727	-2.087	-1.492	-.721	-.246	.578	1.456	1.826	2.787
	-2.659	-2.327	-1.250	-.869	.090	.590	1.179	1.916	2.933
	-2.927	-1.852	-1.257	-.488	-.231	.994	1.618	2.183	2.950
	-2.537	-1.969	-1.193	-.361	-.187	.564	1.566	1.695	2.805
	-2.588	-1.727	-1.480	-.856	.171	.337	1.383	2.116	2.887
	-2.924	-1.930	-1.305	-.581	-.282	.930	1.019	1.678	2.662
	-2.799	-1.795	-1.255	-.403	.217	.533	1.329	1.876	2.684
	-2.896	-2.294	-1.340	-.774	-.280	.935	1.436	2.116	2.586
	-2.639	-2.141	-1.654	-.385	-.175	.828	1.372	1.715	2.422
	-2.875	-1.805	-1.050	-.857	-.262	.495	1.415	1.920	2.662
	-2.525	-2.178	-1.109	-.717	.232	.469	1.431	1.884	2.857
	-2.607	-1.887	-1.508	-.747	-.312	.866	1.533	2.013	2.974
	-2.878	-1.990	-1.353	-.646	.036	.712	1.423	1.963	2.732
	-2.511	-1.953	-1.146	-.359	.149	.625	1.048	1.975	2.935
	-2.775	-1.660	-1.478	-.758	-.187	.804	1.265	2.205	2.800
	-2.338	-2.086	-1.037	-.951	-.281	.358	1.198	2.057	2.629
	-2.635	-1.943	-1.237	-.969	.143	.659	1.536	2.190	2.476
	-2.504	-1.702	-1.593	-.427	.293	.651	1.239	1.841	2.810
	-2.877	-2.252	-1.043	-.527	.282	.866	1.492	2.170	2.513
	-2.760	-2.223	-1.247	-.714	-.189	.750	1.089	2.247	2.892
	-2.581	-2.001	-1.307	-.686	.148	.669	1.050	1.726	2.995
	-2.853	-1.724	-1.070	-.390	.169	.572	1.326	2.188	2.958
	-2.833	-1.944	-1.123	-.694	.318	.656	1.450	1.780	2.665
	-2.533	-2.297	-1.477	-.709	-.035	.702	1.207	1.914	2.421
	-2.715	-2.164	-1.563	-.466	.049	.742	1.287	1.911	2.948
	-2.888	-2.017	-1.558	-.918	-.224	.388	1.309	1.677	2.707
	-2.643	-2.076	-1.060	-.588	-.080	.844	1.140	1.703	2.587
Mean	-2.687	-1.970	-1.315	-.650	.001	.681	1.314	1.974	2.721
S.D.	.168	.185	.194	.195	.202	.179	.184	.191	.204

Table B-4
Difficulties (b), by Stratum, of the Hypothetical
Stradaptive Item Pool Having Item Discriminations of a =2.0

	Stratum								
	1	2	3	4	5	6	7	8	9
	-2.810	-1.760	-1.469	-.528	-.129	.421	1.546	1.999	2.630
	-2.953	-2.054	-1.097	-.630	.137	.747	1.570	2.069	2.898
	-2.773	-1.974	-1.092	-.420	-.321	.453	1.248	1.999	2.334
	-2.698	-2.224	-1.425	-.409	.303	.955	1.490	2.314	2.668
	-2.573	-1.744	-1.155	-.526	-.230	.899	1.113	1.808	2.693
	-2.993	-1.739	-1.077	-.780	-.279	.864	1.530	2.161	2.966
	-2.630	-2.324	-1.078	-.596	-.267	.670	1.391	2.003	2.450
	-2.916	-1.856	-1.456	-.821	-.041	.726	1.272	2.189	2.537
	-2.691	-2.293	-1.604	-.377	-.187	.687	1.665	1.843	2.359
	-2.886	-1.822	-1.325	-.951	-.055	.793	1.637	2.315	2.805
	-2.507	-2.235	-1.496	-.960	.260	.531	1.514	1.905	2.503
	-2.653	-2.222	-1.426	-.918	-.075	.644	1.536	1.797	2.512
	-2.496	-2.067	-1.496	-.752	.001	.453	1.190	2.105	2.967
	-2.633	-2.303	-1.181	-.472	.148	.533	1.220	1.871	2.413
	-2.406	-1.706	-1.027	-.843	-.057	.705	1.615	1.703	2.493
	-2.577	-1.981	-1.334	-.702	.200	.709	1.611	1.796	2.591
	-2.656	-1.737	-1.492	-.981	-.327	.869	1.030	1.935	2.507
	-2.904	-1.842	-1.306	-.685	-.330	.755	1.620	2.149	2.460
	-2.995	-2.077	-1.652	-.914	.099	.524	1.026	2.058	2.421
	-2.691	-1.802	-1.180	-.841	.125	.921	1.498	2.187	2.775
	-2.502	-1.971	-1.590	-.392	.149	.593	1.184	1.991	2.989
	-2.357	-1.778	-1.038	-.919	-.168	.788	1.496	1.829	2.948
	-2.936	-2.008	-1.390	-.504	.070	.657	1.085	2.310	2.955
	-2.992	-1.686	-1.624	-.666	-.301	.608	1.196	2.311	2.393
	-2.704	-2.102	-1.132	-.998	.110	.967	1.652	2.274	2.403
	-2.334	-2.113	-1.039	-.503	-.027	.633	1.519	2.230	2.535
	-2.897	-2.169	-1.560	-.952	-.144	.434	1.124	1.799	2.757
	-2.812	-2.303	-1.346	-.848	.242	.443	1.376	2.166	2.787
	-2.593	-2.127	-1.437	-.786	.215	.649	1.093	2.124	2.760
	-2.835	-2.175	-1.308	-.779	.094	.959	1.019	2.034	2.562
	-2.879	-1.891	-1.504	-.919	.115	.525	1.141	1.845	2.496
	-2.821	-2.313	-1.006	-.715	-.225	.424	1.368	1.752	2.615
	-2.837	-2.048	-1.185	-.428	-.196	.981	1.519	2.200	2.719
	-2.519	-1.965	-1.184	-.405	.194	.427	1.335	1.899	2.358
	-2.883	-1.844	-1.221	-.404	.189	.930	1.138	1.787	2.842
	-2.348	-1.716	-1.265	-.938	.277	.551	1.171	1.731	2.547
	-2.767	-1.864	-1.036	-.594	.198	.622	1.589	1.844	2.950
	-2.726	-2.153	-1.329	-.722	.273	.377	1.127	1.826	2.595
	-2.907	-2.026	-1.372	-.747	.078	.979	1.114	2.287	2.455
	-2.733	-1.851	-1.461	-.847	-.232	.826	1.455	2.102	2.584
	-2.383	-2.246	-1.650	-.613	.076	.912	1.303	1.714	2.783
	-2.787	-2.068	-1.083	-.391	-.329	.808	1.066	1.815	2.600
	-2.952	-1.769	-1.295	-.609	-.219	.672	1.069	2.023	2.676
	-2.637	-2.050	-1.596	-.442	.229	.576	1.552	2.204	2.726
	-2.354	-1.852	-1.048	-.485	.099	.511	1.332	1.812	2.888
	-2.701	-1.712	-1.104	-.750	-.291	.381	1.152	1.866	2.376
Mean	-2.710	-1.990	-1.308	-.684	-.012	.676	1.337	2.000	2.637
S.D.	.192	.196	.198	.196	.202	.183	.208	.189	.193

Table B-5
Item Difficulties (b) for Hypothetical Conventional Tests, at Three Discrimination Levels

Discrimination (a)		
0.5	1.0	2.0
-.052	.100	.273
.021	.070	.041
-.027	-.246	-.275
-.163	-.155	.305
-.110	-.040	.038
-.001	-.034	-.261
.277	-.025	.140
-.236	-.310	-.090
-.270	-.283	.277
-.290	.254	-.254
.147	-.237	.038
-.247	.256	.154
.126	.031	-.329
-.188	.274	.062
.245	-.222	.082
-.196	.225	-.282
-.240	.242	.199
.307	.026	.028
-.091	.020	.269
-.083	.255	-.079
-.050	-.311	-.119
-.127	.169	.303
-.200	-.172	-.059
.092	-.313	.103
.037	-.306	.037
-.169	.045	.228
-.082	-.136	-.243
.328	-.159	.265
.021	.255	.008
.181	.313	-.198
.263	.126	-.282
-.033	.057	-.254
-.099	.054	-.206
.263	.312	-.315
.174	-.045	-.177
.327	-.325	-.133
-.321	-.251	.053
.077	.043	-.122
.313	.011	.073
.089	.014	-.033
.014	.001	.128
.282	.166	-.280
.018	.294	-.194
-.087	-.244	.046
.245	.150	.016
.220	.021	-.128
-.049	.144	-.316
-.201	.219	.105
-.287	.195	-.281
-.111	.268	-.328
.110	.221	-.293
-.253	.085	.143
-.262	.138	-.104
-.050	.247	-.217
-.164	-.136	-.126
-.243	.012	.160
.320	-.022	-.112
-.257	-.043	.202
-.211	-.265	.185
-.172	.001	.157
Mean Difficulty -.019	.017	-.033
S.D. Difficulty .194	.191	.193

APPENDIX C

Computer Hardware and Software Used for the Simulations

The Computer

The computer used for this research was a Hewlett-Packard 9600E real-time computer system, which is based on a 2100S central processer with a memory consisting of 32K of 16-bit words. Peripheral equipment available consisted of one disk with a capacity of about 5 million ASCII characters, a high speed paper tape reader, a teletype and four cathode ray terminals (CRTs).

Integer numbers were represented in a single computer word with a maximum value of ±32768. Real numbers were represented as two 16-bit words having about six significant digits. For long addition where this was not sufficient precision, double-precision arithmetic was used providing 13-digit significance by using three 16-bit words.

Approximately half of the total memory was used by the computer operating system, but the half remaining proved adequate for all simulation programs used. The disk had room to store test scores for 15,000 testees after space was provided for system programs and simulation programs.

The Program System

To make efficient use of available computer time, it was necessary to run the simulation programs from 5:00 each evening until 8:00 the next morning as well as from 5:00 Friday evening until 8:00 Monday morning. Since the simulation program could fill its available storage space in an hour or so, it was obvious that the data would have to be generated, analyzed, erased, and regenerated. To make this process semi-automatic, so that a programmer would not have to be present to start a new program each hour, a program system organized as shown in Figure C-1 was constructed.

Scheduling. The first step in running the simulation system was to write a schedule of programs and enter this through the teletype. This schedule was then read by the scheduling program which in turn scheduled a certain processing program to run. The schedule might have been, for example, as follows: 1) generate and score 1000 conventional test response records using item parameter file 1 and a normal ability distribution; 2) run a correlation program to correlate the generating ability with the test score; 3) print the results; 4) generate 15,000 test records using a rectangular ability distribution; 5) calculate the information values; 6) print the results; and 7) stop.

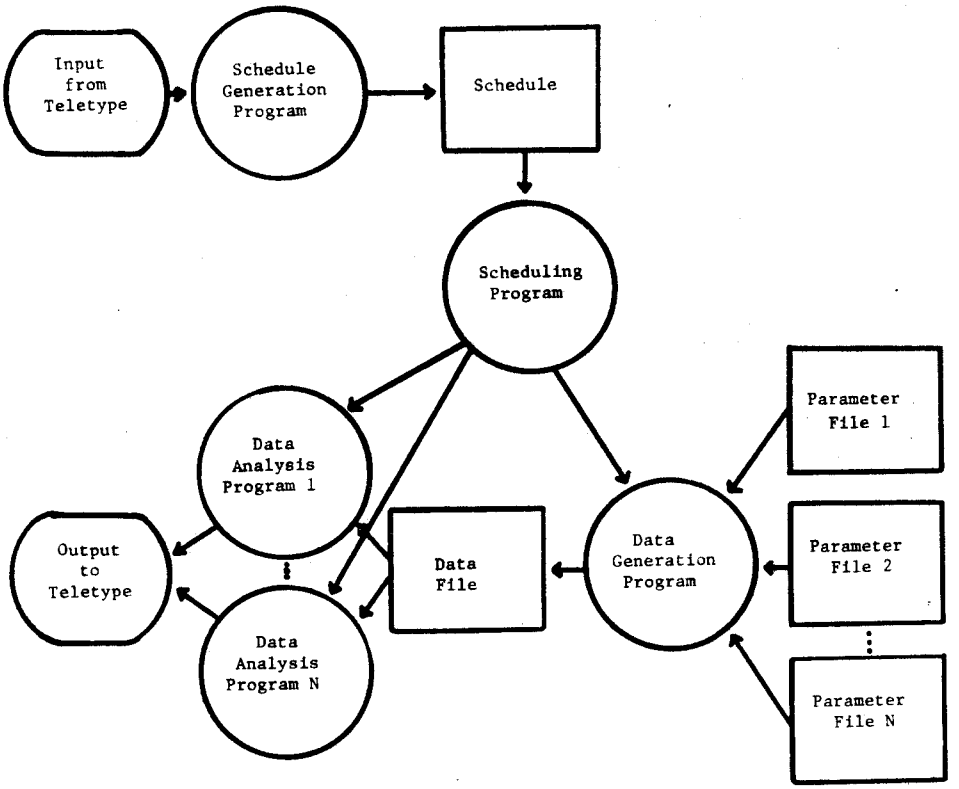
The scheduling program would read the first element of the schedule and, seeing that it said to generate data, would schedule the data generating program 1000 times, etc. After each program finished, control was returned to the scheduling program which then scheduled another process or stopped.

The data generation program. The core of this system was the data generation program. On orders from the scheduling program, it first selected an item parameter file to work from and then generated an ability either randomly from a normal distribution or at a fixed level as dictated by the schedule. If the schedule called for a stradaptive test, an initial ability estimate was generated which was either fixed at the mean of the ability distribution (so

everyone entered at the middle stratum) or sampled from a normal distribution about the mean of the ability distribution with a varying degree of correlation with the generating ability. A test protocol was then generated and scored with length being controlled either by the standard termination criterion for Variable-Length Stradaptive, or a schedule-dictated termination criterion for Fixed-Length Stradaptive. The scores from this test were then written on the data file and the procedure was repeated until a sufficient number of testees had been run.

Figure C-1

Schematic Representation of the Simulation Program System (Arrows Represent Flow of Information)



The data analysis programs. The data analysis programs read from the data file generated by the data generation program, calculated appropriate statistics, and printed the results on the teletype. Following this, the data file was free to be written upon again. All statistical analyses were programmed specifically for this research, using common formulas for descriptive statistics and correlations, and formulas described in the Data Analysis section for informational statistics.

REFERENCES

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).

Betz, N.E. & Weiss, D.J. Empirical and simulation studies of flexilevel ability testing. Research Report 75-3, Psychometric Methods Program, Department of Psychology, University of Minnesota, 1975.

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

Ghiselli, E.E. Differentiation of individuals in terms of their predictability. Journal of Applied Psychology, 1956, 40, 374-377.

Guilford, J.P. Fundamental Statistics in Psychology and Education (2nd Ed.). New York: McGraw-Hill, 1950.

Gulliksen, H. Theory of Mental Tests. New York: Wiley, 1950.

Jensema, C.J. An application of latent trait mental test theory to the Washington Pre-College Testing Battery. Unpublished doctoral dissertation, University of Washington, 1972.

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).

Loevinger, J. The attenuation paradox in test theory. Psychological Bulletin, 1954, 51, 493-504.

Lord, F.M. & Novick, M.R. Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1968.

Lord, F.M. Some test theory for tailored testing. In W.H. Holtzman (Ed.), Computer-Assisted Instruction, Testing, and Guidance. New York: Harper and Row, 1970.

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).

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

Sitgreaves, R. A statistical formulation of the attenuation paradox in test theory. In H. Solomon (Ed.), Studies in Item Analysis and Prediction. Stanford: Stanford University Press, 1961.

Sympson, J.B. Problem: evaluating the results of computerized adaptive testing. In D.J. Weiss (Ed.), Computerized adaptive trait measurement: problems and prospects. Research Report 75-5, Psychometric Methods Program, Department of Psychology, University of Minnesota, 1975.

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

Urry, V.W. Computer-assisted testing: the calibration and evaluation of the verbal ability bank. Technical Study 74-3. Research Section, Personnel Research and Development Center, U.S. Civil Service Commission. Washington, D.C., 1974.

Vale, C.D. Problem: strategies of branching through an item pool. In D.J. Weiss (Ed.), Computerized adaptive trait measurement: problems and prospects. Research Report 75-5, Psychometric Methods Program, Department of Psychology, University of Minnesota, 1975.

Vale, C.D. & Weiss D.J. A study of computer-administered stradaptive ability testing. Research Report 75-4, Psychometric Methods Program, Department of Psychology, University of Minnesota, 1975.

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. 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).