DOCUMENT RESUME

ED 118 604 TM 005 090

AUTHOR	Waters, Brian K.
TITLE	Empirical Investigation of the Stradaptive Testing Model for the Measurement of Human Ability.
INSTITUTION	Air Force Human Resources Lab., Williams AFB, Ariz. Flying Training Div.
SPONS AGENCY	Air Force Human Resources Lab., Brooks AFB, Texas.
REPORT NO	AFHRL-TR-75-27
PUB DATE	Oct 75
NOTE	72p.
EDRS PRICE	MF-$0.83 HC-$3.50 Plus Postage
DESCRIPTORS	*Ability; Branching; Comparative Analysis; *Computer Oriented Programs; Group Tests; Individual Differences; Item Analysis; Models; Response Style (Tests); Scoring; *Testing; Test Reliability; *Test Validity; Verbal Ability
IDENTIFIERS	*Stradaptive Testing

ABSTRACT

This study empirically investigated the validity and utility of the stratified adaptive computerized testing model (stradaptive]developed by Weiss (1973). The model presents a tailored testing strategy based on Binet IQ measurement theory and Lord's (1972) modern test theory. Nationally normed School and College Ability Test Verbal analogy items (SCAT-V) were used to construct an item pool. Item difficulty and discrimination indices were rescaled to normal ogive parameters on 249 items. Freshmen volunteers at Florida State University were randomly assigned to stradaptive or conventional test groups. Both groups were tested via cathode-ray-tube (CRT) terminals coupled to a Control Data Corporation 6500 computer. The conventional subjects took a SCAT-V test, while the stradaptive group took individually tailored tests drawn from the same item pool. Results showed significantly higher reliability for the stradaptive group, and equivalent validity indices between stradaptive and conventional groups. Three stradaptive testing strategies averaged 19.2, 26.5, and 31.5 items per subject as compared with 48.4 items per conventional subject. A 50% reduction from conventional test length produced an equal precision of measurement for stradaptive subjects. Item latency comparisons showed the stradaptive group required significantly longer per item than conventional group members. It is recommended that time rather then number of items be used in future adaptive research as a dependent variable. (Author/DEP)

AFHRL-TR-75-27

TM

AIR FORCE HUMAN RESOURCES LABORATORY

EMPIRICAL INVESTIGATION OF THE STRADAPTIVE TESTING MODEL FOR THE MEASUREMENT OF HUMAN ABILITY

By

Brian K. Waters, Major, USAF

FLYING TRAINING DIVISION
Williams Air Force Base, Arizona 85224

October 1975

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235

NOTICE

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This disseration was submitted by Flying Training Division, Air Force Human Resources Laboratory, Williams Air Force Base, Arizona 85224, under project 1121, with Hq Air Force Human Resources Laboratory (AFSC), Brooks Air Force Base, Texas 78235.

The views expressed are those of the author and do not necessarily reflect the views of the United States Air Force or the Department of Defense.

This report has been reviewed and cleared for open publication and/or public release by the appropriate Office of Information (OI) in accordance with AFR 190-17 and DoDD 5230.9. There is no objection to unlimited distribution of this report to the public at large, or by DDC to the National Technical Information Service (NTIS).

This technical report has been reviewed and is approved.

WILLIAM V. HAGIN, Technical Director
Flying Training Division

Approved for publication.

HAROLD E. FISCHER, Colonel, USAF
Commander

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFHRL-TR-75-27
2. GOVT ACCESSION NO.:
3. RECIPIENT'S CATALOG NUMBER:

4. TITLE (and Subtitle): EMPIRICAL INVESTIGATION OF THE STRADAPTIVE TESTING MODEL FOR THE MEASUREMENT OF HUMAN ABILITY

5. TYPE OF REPORT & PERIOD COVERED: Dissertation

6. PERFORMING ORG. REPORT NUMBER:

7. AUTHOR(s): Brian K. Waters

8. CONTRACT OR GRANT NUMBER(s):

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Flying Training Division
Air Force Human Resources Laboratory
Williams Air Force Base, Arizona 85224

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
62703F
11210310

11. CONTROLLING OFFICE NAME AND ADDRESS
Hq Air Force Human Resources Laboratory (AFSC)
Brooks Air Force Base, Texas 78235

12. REPORT DATE: October 1975

13. NUMBER OF PAGES: 70

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

15. SECURITY CLASS. (of this report): Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

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

18. SUPPLEMENTARY NOTES

The views expressed are those of the author and do not necessarily reflect the views of the United States or the Department of Defense.

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

adaptive testing
tailored testing
program testing
computer-based testing

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

This study empirically investigated the validity and utility of the stratified adaptive computerized testing model (stradaptive) developed by Weiss (1973). The model presents a tailored testing strategy based upon Binet IQ measurement theory and Lord's (1972) modern test theory.

Nationally normed School and College Ability Test Verbal analogy items (SCAT-V) were used to construct an item pool. Item difficulty and discrimination indices were rescaled to normal ogive parameters on 249 items.

One hundred and three freshmen volunteers at Florida State University were randomly assigned to stradaptive or conventional test groups. Both groups were tested via cathode-ray-tube (CRT) terminals coupled to a Control Data Corporation 6500 computer.

Item 20 (Continued)

The conventional subjects took a SCAT-V test essentially as published, while the stradaptive group took individually tailored tests drawn from the same item pool.

Results showed significantly higher reliability for the stradaptive group, and equivalent validity indices between stradaptive and conventional groups. Both KR-20 and parallel-forms reliabilities were computed for the stradaptive group.

Three stradaptive testing strategies averaged 19.2, 26.5 and 31.5 items per subject as compared with 48.4 items per conventional subject. A 50% reduction from conventional test length produced an equal precision of measurement for stradaptive subjects.

Item latency comparisons showed the stradaptive group required significantly longer per item (about 11%) than conventional group members. The author recommended that time rather than number of items be used in future adaptive research as a dependent variable.

Further investigation of the stradaptive model was recommended with comparisons between variable and fixed test termination rules.

TABLE OF CONTENTS

		Page
I.	Introduction	5
II.	Review of Related Research	6
	Fixed Number of Stages	10
	Constant Step Sizes	10
	Variable Step Sizes	14
	Variable Number of Stages	16
	Constant Step Size	16
	Variable Step Sizes	17
	Summary of the Literature on Adaptive Testing	18
III.	The Stradaptive Testing Model	19
	The Item Bank	20
	Item Content and Format	23
	Computer Program for Model Implementation	24
	Instructional Sequence	24
	Testing Sequence	24
	Scoring Methods	25
	Termination Rules	26
	Stradaptive Test Output	27
IV.	Procedures	27
	Item Pool Construction	27
	Subject Pool	30
	Research Design	31
	Data Collection	31
	Data Analysis	32
	Attitudinal Data	32
V.	Results and Discussion	33
	Linear Test Reliability	34
	Linear Test Validity	35
	Stradaptive Pool Item Stratification	35
	Stradaptive Total-Test Reliability	36
	Stradaptive Test Validity	38
	Linear vs. Stradaptive Comparisons	40
	Omitting Tendency	41
	Correlation Between Scores of Subjects Who Took Both Stradaptive and Linear Tests	41
	Attitudinal Information	42
	Testing Costs	42
V.	Conclusions and Implications for Future Research	42
	References	45

Table of Contents *(Continued)*

Appendix A: Item Statistic Comparison 50

Appendix B: Transformation of Traditional Item Difficulty (P_g) and Biserial Correlation (r_g') to Normal Ogive Parameters b_g and a_g 57

Appendix C: Form Letter 64

Appendix D: Description of Data 66

LIST OF ILLUSTRATIONS

Figure		Page
1	Adaptive testing research strategies	10
2	Efficiency of measurement as a function of ability level (after Lord, 1970; 1971a, b, c)	11
3	Bayroff's example two stage adaptive test (1964)	12
4	Example of 8-step pyramidal adaptive test. (From Bayroff, 1964)	13
5	Distribution of items, by difficulty level, in a stradaptive test (from Weiss, 1973)	21
6	Scatterplot of relationship between A_g and B_g	23
7	Entry point question for determining subject ability estimate (from Weiss, 1973)	24
8	Example of stradaptive testing report	28
9	Research design for linear versus stradaptive group assignment and comparison	32

LIST OF TABLES

Table		Page
1	Alternate Terminology Used to Describe Adaptive Testing Strategies and Their References	7
2	Classification of Research Studies on Adaptive Testing by Type and Branching Strategy	9
3	Item Difficulties (B) and Discriminations (A), based on Normal Ogive Parameter Estimates, for the Stradaptive Test Item Pool	22
4	Comparison of SCAT Series II Verbal Forms 1A, 1B, 1C, 2A, & 2B (N = 3,133)	29
5	Descriptive Statistics of Difficulty (b_g) and Discrimination (a_g) Normal Ogive Parameter	30
6	Comparison of Florida 12th Grade Verbal Test Scores (1973 Statewide Administration vs. Subject Sample)	30
7	Comparison of Distributions of Linear and Stradaptive Group Florida 12th Grade Verbal Scores	31
8	Comparison of Distributions of 5 Linear Subtests	33
9	Distribution of Pooled Linear Test Scores	33
10	Analysis of Variance for Linear Test Person by Item Matrix	34

List of Tables *(Continued)*

11 Comparison of Difficulty Distributions (P_c) for Linear and Stradaptive Groups 35

12 Reported Correlations of SCAT-V Scores with External Criteria 35

13 Proportion of Items in Each Stratum Actually used in CRT Stradaptive Testing (N = 55) 35

14 Analysis of Variance of Scoring Method 8 of Stradaptive Test Person-by-Item Matrix 36

15 Comparison of Parallel-Forms Reliabilities for 10 Stradaptive Test Scoring Methods under Three Termination Rules Stepped-Up to 50 Items 37

16 Comparison of Scoring Method 8 Parallel Form Reliability with KR-20 Reliability Over Three Termination Rules Stepped Up to 50 Items 38

17 Comparison of Validity Coefficients of 10 Stradaptive Test Scoring Methods Under Three Termination Rules 38

18 Effect of the Four Most Reliable Stradaptive Scoring Methods Correlation with 12V, Corrected for Attenuation 39

19 Comparison of Linear Test with Scoring Method 8 Under Three Termination Rules of the Stradaptive Test 40

20 Comparison of Average Number of Items for Linear Test and Three Termination Methods of Alternate Form Stradaptive Tests 40

21 Comparison of Distributions of Item Latency Between Linear and Stradaptive Groups 41

22 Linear and Stradaptive Scores of Subjects Who Took Both Tests 41

EMPIRICAL INVESTIGATION OF THE STRADAPTIVE TESTING MODEL FOR THE MEASUREMENT OF HUMAN ABILITY

I. INTRODUCTION

This study investigated the validity and utility of the stratified adaptive, "stradaptive" computerized testing model proposed by Weiss and his colleagues in the Psychometric Methods Program, University of Minnesota. The stradaptive model, theoretically, could provide a highly efficient means of assessing ability in large-scale testing situations. Such a model could readily be implemented in military training or industrial selection and classification situations.

The model is based upon the early work of Binet in the measurement of intelligence and upon Lord's recent theoretical research in tailored testing. The model also utilizes modern latent trait theory and parameter estimates as detailed in Lord and Novick (1968).

Weiss and his associates have reported the theoretical development of the stradaptive model (Weiss, 1973; DeWitt & Weiss, 1974; McBride & Weiss, 1974) including some examples of individual results. To date, no full empirical studies of the model have been published. Weiss' exploratory evidence appears promising, but leaves many questions unanswered. He suggests ten possible scoring methods, yet offers no evidence as to the "best" method. The evaluation of scoring methods appropriate for tailored testing was one of the secondary goals of this study. The primary goal of this study was the validation of the model itself.

Comparisons were made between the stradaptive group test scores and conventional group test scores, both presented via a cathode-ray-tube mode of testing. Reliability and validity indices relative to the specific subject sample used in this experiment were calculated.

The stradaptive model is very sensitive to the accuracy of item parameter estimates. In order to minimize item parameter estimation errors, a large norming group is essential. Weiss and his colleagues were well aware of this constraint, and have suggested specific procedures for establishing a reliable item pool for adaptive testing (Larkin & Weiss, 1974). Nevertheless, the item pool used in their reported examples of stradaptive testing were based on item parameter estimates calculated from norming groups of less than 200 subjects. In this current study, items from the School & College Ability Test (SCAT) Series II Verbal Ability test (1966) which had been nationally normed on a group of 3133 examinees comparable to the subjects in this experiment were used. These items should provide more trustworthy item parameters for use in the investigation of the model.

Determining the merits of a particular testing strategy has been a major problem in previous studies of tailored testing. In any kind of tailored test, different examinees take different test items, thus prohibiting many classical measurement indices of "goodness." Reliability assessment, particularly, has suffered due to this problem. Traditional internal consistency calculations are not possible, and procedures such as Hoyt's (1941) ANOVA reliability estimate apparently have unacceptable underlying assumptions (such as item independence when applied to tailored testing). One goal of this study was to determine an alternate form reliability of the stradaptive test scores and to compare this index with a Hoyt-type reliability index. This alternate form reliability index would provide a measure of the "goodness" of the stradaptive model as well as of the ANOVA reliability estimation procedure.

Validity, as well as reliability, must be adequate for a testing strategy to be "good." Eighty-seven of the 103 subjects in this experiment had previously taken the Florida 12th Grade Verbal test composed of items identical in form to the SCAT Series II Verbal items and 12 subjects had 12th Grade Verbal score estimates derived from American College Testing (ACT) or College Entrance Examination Board (CEEB) Verbal Test scores. Both the Florida 12th Grade test and the SCAT tests were produced by Educational Testing Service (ETS) and purportedly measured the same psychological dimension. Like the SCAT, the Florida 12th Grade was normed on a large sample of subjects comparable to the subjects in this experiment. Thus, the 12th Grade scores provided ideal external criteria scores for the stradaptive validity examination.

Item latency data was collected on all subjects in this experiment. Since each item was tailored to the examinee's ability level, it was hypothesized that examinees on a tailored test would take more time per item than on a conventional test. If this hypothesis were supported, the dimension of testing time must be considered in evaluating a tailored testing model.

There is little doubt that the use of interactive computer testing will increase enormously in the coming decade. Research in this area has just started to reveal some of the potential benefits of tailored testing to institutions and individuals alike. Improved measurement accuracy and efficiency through the use of some kind of adaptive, computer-based testing, appear to be among these potential benefits. This study empirically investigated one such proposal, the stradaptive testing model.

II. REVIEW OF RELATED RESEARCH

As the term implies, adaptive testing is defined as a method of test construction wherein the items presented to a specific subject are selected iteratively dependent upon his previous responses, thus "adapting" the test to the subject. Many terms have been used in the literature to refer to such an item selection strategy (Table 1). In this paper, the comprehensive term "adaptive testing" will be used to include any or all of the testing strategies listed in Table 1.

Adaptive testing had its beginnings in the early work of Binet on the measurement of intelligence. The original Binet scale and the current version, the 1960 Stanford-Binet Scales (Terman & Merrill, 1960) utilized an adaptive strategy to estimate a subject's IQ. The testing begins with the examiner selecting the first item to be presented, based upon his judgment of the subject's ability level. Once testing starts, the examiner may present the items in varying orders, based somewhat upon examinee responses. The basal and ceiling ages of the subject are estimated in order to present items which are neither too easy nor too hard for the subject. This is done through the construction of groups of items whose difficulties are centered around "mental ages," that is, "peaked" tests are formed in which about 50% of the norming group of that chronological age responded with a correct answer to those items. Thus, the Stanford-Binet can be looked upon as a series of mini-tests designed to provide an efficient measure of the ability of each subject.

Theoretically, individual testing, as in the case of the Binet, should provide more accurate measurement than group testing. Nevertheless, individual testing strategies do have weaknesses. Obviously, the major problem is the cost of administration. These tests must be administered by a highly-trained examiner working on a one-to-one basis with the subject. Such expenditure may be warranted for an individual case basis when subjects are referred through external evaluations, but are clearly impractical on any large scale.

In addition to the cost deterrent, individual testing is plagued by several more technical problems. Weiss and Betz (1973) cite numerous research studies suggesting differential examiner effects. Differential scoring effects were cited, as well as interaction effects between the personality and social attributes of both examiners and examinees. Thus, the theoretical gains in measurement efficiency attributed to an individual testing strategy may well be offset by the added variance in test scores due to uncontrolled factors in the testing process.

The paper and pencil mode of item presentation is, of course, the most common testing strategy. An enormous volume of theoretical and empirical work has been done under the banner of classical measurement theory. This field has made giant strides through the reduction of measurement error and thus, the improved utility of the scales. Many practical situations demand that all subjects must take the same collection of test items, with identical time limits, via the paper and pencil mode of presentation. Nevertheless, it must be realized that certain limitations are inherent in conventional test administration.

Careful training and standardization of group test administrators is intended to control for many of the inadequacies of individual test administration. Research evidence exists which shows that uncontrolled examiner variables are still present. Weiss and Betz (1973) extensively discussed five major areas in which unwanted variance enters the group measurement process:

1. Administrator variables, such as sex or race:
2. Answer sheet effects, in which answer sheet formats differentially affect test performance:
3. Item arrangement effects within a test:
4. Timing and time limit effects:

and

5. An effect resulting from the standardized set of items which is administered to all examinees.

Stanley (1971) suggests that the effective length of a test is considerably shorter than the actual length of the test for a specific examinee, since many items are too easy and many are too hard. The easy items are a

TABLE 1

Alternate Terminology Used to Describe Adaptive Testing Strategies and Their References

TESTING STRATEGY	REFERENCES
ADAPTIVE	Kappauf, 1969; Wood, 1971; Wood, 1972; Betz & Weiss, 1973; Weiss, 1973; Weiss & Betz, 1973; DeWitt & Weiss, 1974; Larkin & Weiss, 1974; McBride & Weiss, 1974
BAYESIAN	Novick, 1969; Owen, 1969, Urry, 1970; Urry, 1971
BRANCHING	Waters, 1964; Bayroff, 1969; Waters, 1970; Bayroff, 1971; Waters & Bayroff, 1971, Bryson, 1971
FLEXILEVEL	Lord, 1971b, d; Olivier, 1973; Olivier, 1974
MULTI-LEVEL	Angoff & Huddleston, 1958
PROGRAMMED	Bayroff, 1964; Hubbard, 1966; Bayroff & Seeley, 1967; Cleary, Linn & Rock, 1968a,b; Linn, Rock & Cleary, 1969
RESPONSE-CONTINGENT	Wood, 1973
SEQUENTIAL	Cowden, 1946; Wald, 1946; Moonan, 1950; Krathwohl & Huyser, 1956; Bayroff, Thomas & Anderson, 1960; Paterson, 1962; Seeley, Morton & Anderson, 1962; Cronbach & Gleser, 1965; Hansen, 1969; Kappauf, 1969; Linn, Rock & Cleary, 1970; Wood, 1971; Wood, 1972
TAILORED	Lord, 1968; Owen, 1969; Owen, 1970, Stocking, 1969; Wood, 1969; Green, 1970; Holtzman, 1970; Lord, 1970; Lord, 1971a,c,e; Kalisch, 1974

waste of time and testing costs, while the too hard items encourage guessing and add all the measurement problems associated with this source of extraneous variance. Thus, a standard set of items, peaked at the mean of the norming group is only truly optimal for a subject of mean ability on the dimension being measured. Consistent with this, information theory research has shown that a test peaked at a difficulty value of .5 provides optimum measurement (maximizes internal consistency) *for examinees of the subject's ability level* (Hick, 1951; Lord, 1970, 1971, 1971a, 1971d, 1971e).

In addition to the previously mentioned problem wherein the standard set of items contributes to guessing, another serious problem arises. Many research studies have shown that guessing is not a consistent trait throughout the ability continuum (Lord, 1957, 1959; Baker, 1964; Nunnally, 1967; Boldt, 1968). Low ability subjects guess more often than high ability subjects, creating differential measurement accuracy along the ability continuum.

The literature implies that both conventional paper and pencil group tests, and traditional, individually administered tests are not always optimally suited to large-scale ability testing. Adaptive testing appears to offer a feasible and practical alternative to these two modes of test administration. It involves selecting a test item for presentation based upon the subject's response to the previous item or items.

The principle underlying the Binet testing strategy—e.g., that the difficulty of the test items selected for a given subject should be peaked around the subject's ability level, not the total group's ability level, is also the basis of the stradaptive model.

Considerable research has been done in the last twenty years to find a method of testing which will accomplish this goal. Figure 1 depicts a three dimensional (3 x 2 x 2) model of adaptive testing research strategies categorized according to (1) type of research (empirical, simulated or theoretical); (2) whether the number of items (or stages) is fixed for all examinees; and (3) whether the item difficulty step-size between stages is fixed or variable throughout the test.

Table 2 lists the particular cells of Figure 1 with research studies reviewed noted in the appropriate cells. It is hoped that Table 2 will provide a helpful reference to the literature for future researchers concerned with adaptive testing. The balance of this literature review will refer to Table 2 and discuss research results cell-by-cell.

Any classification system such as that used in Table 2 and Figure 1 require many arbitrary categorization decisions. For the purposes of this paper, an *empirical study* was defined as one in which "real-live" subjects provided the source of the data in a research study. Studies in which existing data banks were reanalyzed "as if" the subjects had proceeded through the test according to some other strategy than they actually did were classified as *simulated studies*. Computer-generated monte carlo studies were included in this category. The *theoretical category* included both mathematical and non-mathematical discussions of adaptive testing strategies and provided somewhat of a catchall for research studies that did not seem to fit the other two classifications. Some studies were multiple-classified if comparisons were made between adaptive strategies of more than one type.

The dimension "step-sizes" similarly required some arbitrary assignments. Two-stage testing, for example, is not always structured according to fixed step-sizes, though theoretically, it could be. Nevertheless, this adaptive strategy was considered to be fixed step-size rather than the "true" variable step-size strategies as is the case in the Robbins-Munro technique. A study was assigned to the fixed number of *stages* dimension if all examinees in a comparison group took the same number of items, regardless of the number of stages involved in the branching strategy.

As shown by the left half of Table 2, about two thirds of the adaptive testing papers reviewed were concerned with a fixed number of stages per test. This concentration is understandable. First, having all examinees take the same number of items simplifies statistical analysis immensely, particularly when estimating internal consistency reliability. Stanley (1971) has shown a method for determining this index despite unequal numbers of items per subject, but his paper post-dated much of the reported research in adaptive testing. Secondly, the training of the majority of psychometricians has been under classical measurement theory in which all subjects are completely crossed with all items. Finally, testing large numbers of subjects with tests of different lengths probably had to await the development of computer-based testing technology. This last point is vividly supported by the fact that 13 of the 15 variable number of stage studies reviewed have been published since 1968.

The second dimension in Table 2 "step sizes," like "number of stages" was predominantly concentrated in one classification. Two thirds of the studies reviewed analyzed only constant step sizes. The "constant

TABLE 2

Classification of Research Studies on Adaptive Testing by Type and Branching Strategy

TYPE OF RESEARCH STUDY	BRANCHING STRATEGY: FIXED NUMBER OF STAGES — CONSTANT STEP SIZES	BRANCHING STRATEGY: FIXED NUMBER OF STAGES — VARIABLE STEP SIZES	BRANCHING STRATEGY: VARIABLE NUMBER OF STAGES — CONSTANT STEP SIZES	BRANCHING STRATEGY: VARIABLE NUMBER OF STAGES — VARIABLE STEP SIZES
SIMULATED	Cleary, Linn & Rock (1968a) Cleary, Linn & Rock (1968b) Linn, Rock & Cleary (1970) Waters & Bayroff (1971)	Bryson (1971) Paterson (1962)	Linn, Rock & Cleary (1970)	Kalisch (1974) Urry (1970) Urry (1971)
THEORETICAL	Lord (1970) Lord (1971a) Lord (1971b) Lord (1971c) Lord (1971d) Lord (1971e)	Lord (1970) Lord (1971a) Lord (1971c) Stocking (1969)	Green (1970) Kappauf (1969) Cronbach & Gleser (1965) Wald (1946) Weiss (1973)	Novick (1969) Owen (1969) Owen (1970)
EMPIRICAL	Angoff & Huddleston (1958) Bayroff (1964) Bayroff & Seeley (1967) Betz & Weiss (1973) Hansen (1969) Larkin & Weiss (1974) Krathwohl & Huyser (1956) Olivier (1974) Wood (1971)	Bayroff, Thomas & Anderson (1960) Bryson (1971) Seeley, Morton & Anderson (1962)		Ferguson (1969) Ferguson (1971) Wood (1971)

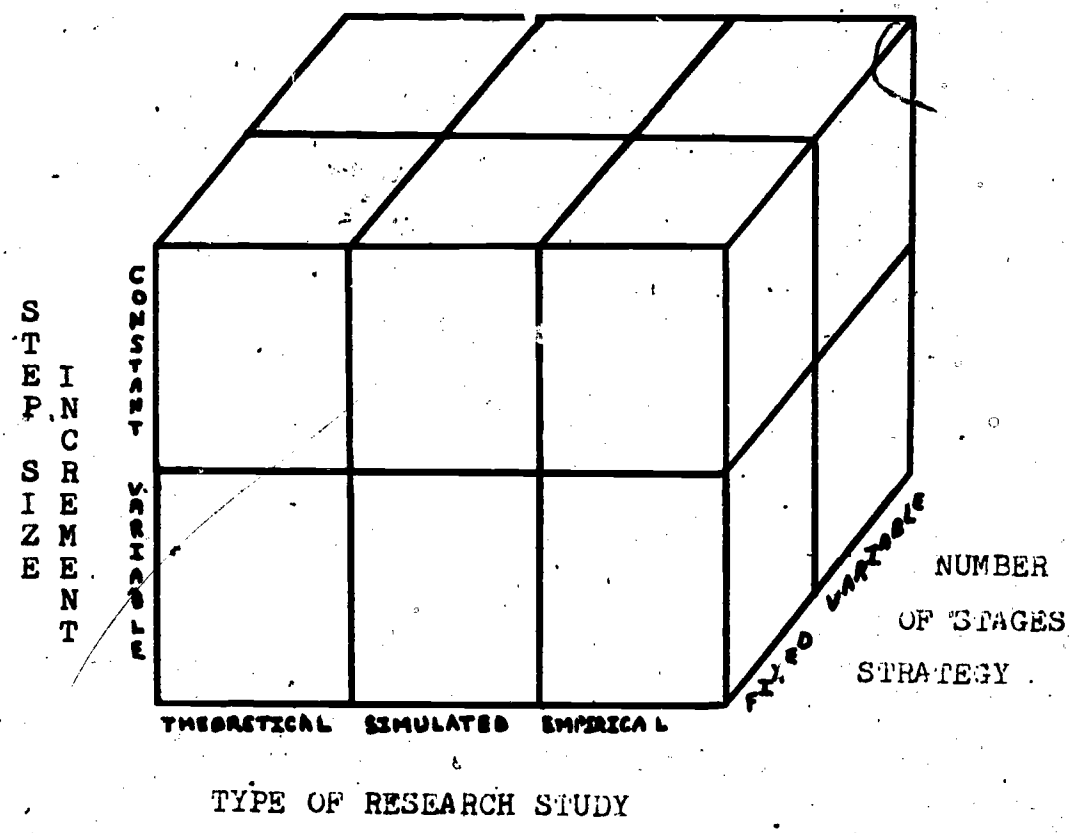


Figure 1. **Adaptive testing research strategies.**

step-size" categorization included both pyramidal and multiple stage tests. In pyramidal testing, items are grouped by difficulties over a set number of stages, while multiple stage tests include routing and measurement stages with a set number of items per stage and a given number of stages for all subjects.

The third dimension of Table 2, "*type of research study*" shows a fairly even distribution among empirical, simulated and theoretical work. One would expect the theoretical papers to precede the empirical model validation studies. However, the three levels of this dimension have been published concurrently throughout the last fifteen years or so.

The balance of this chapter will consider each of the three dimensions of Figure 1 and briefly summarize consistent results within each cell.

Fixed Number of Stages

Constant Step Sizes

Theoretical studies. Lord's six papers (1970, 1971, a,b,c,d,e) investigated the measurement effectiveness of both fixed and variable step size strategies within several varieties of fixed number of stages. His work utilizes the item characteristic curve theory (Lord, 1972) under a specific set of assumptions which will be discussed in Chapter III of this paper. Lord's theoretical analysis of two stage testing (1971c) varied the number of items presented to each subject in the routing and measurement tests, the distributions of items between the two stages and whether guessing was assumed to be present or not. His results were presented in the form of graphic comparisons between the several adaptive testing strategies and a 60 item peaked conventional test using information functions to evaluate the amount of information yielded (Figure 2).

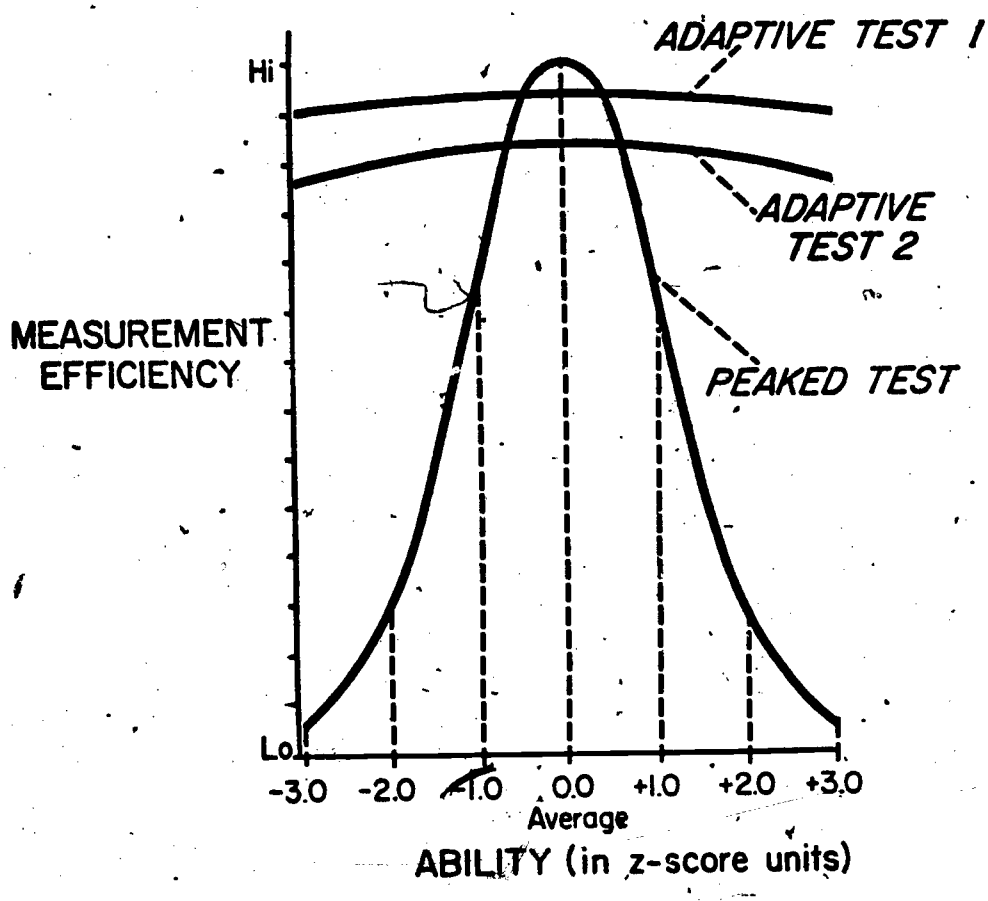


Figure 2. **Efficiency of measurement as a function of ability level (after Lord, 1970; 1971a, b, c).**

He concluded that the best of the two stage strategies provided almost as effective measurement near the mean of the ability continuum, with relatively greater improvement as a subject's ability level departed from the mean ability of the group. He found that guessing decreased the effectiveness of measurement for low-ability subjects, but affected high ability estimates much less.

Lord's theoretical development (1971b) and evaluation (1971d) of flexilevel testing was an attempt to implement the adaptive testing concepts under a paper and pencil mode of test presentation. Lord's analysis compared a 60-item flexilevel test with a 60-item conventional test, both tests with assumed equal item discriminations and a third test peaked at two points along the ability continuum. He found the flexilevel test superior in information provided throughout the range of abilities. As with the two stage testing, the conventional peaked test measured more effectively than the adaptive test in the center of the distribution of scores, but the flexilevel ability estimate was more accurate for at least 30% of the population. Unfortunately, the only empirical study to date of flexilevel testing (Olivier, 1974) found reduced efficiency of measurement throughout the ability continuum.

Simulated studies. Five research studies on simulated data were reviewed. These concentrated upon a fixed number of stages and constant step sizes. Three of these studies were made by Cleary, Linn, and Rock (1968a, 1968b, 1970) using 190 items from SCAT and STEP item banks which were then reanalyzed *as if* the

subjects had proceeded through the item pool in an adaptive fashion. They compared seven strategies of two-stage adaptive testing with 10, 20, 30, 40, and 50 item conventional tests from the same pool. They found one of the adaptive procedures correlated highest with total score, followed by the conventional tests and then the rest of the adaptive tests. The authors estimated an improvement of about 35% over the best short conventional test on a comparable number of items by the best adaptive strategy. Validity coefficients in every case but one showed higher correlations with external criteria for adaptive tests than the conventional tests of equal length.

Waters and Bayroff (1971) used hypothetical 5, 10, and 15 item conventional tests for comparison with 5 and 10 item branching tests, varying item difficulty ranges and the item-biserial index. Their study showed that adaptive tests yielded higher validities than any of the conventional tests for tests made up of items with a biserial index at .60 or .80 and equal validity coefficients at a .40 biserial.

The simulated results of the Cleary group and the Bayroff group were very similar and comparable to the empirical results reported in the following section of this paper.

Empirical studies. The eight empirical research studies reviewed by the author investigated adaptive testing strategies having a fixed number of items or stages and constant step sizes. Two major varieties of adaptive testing have been empirically evaluated; two-stage testing and multi-stage testing. Typically, in the former strategy, a routing test with a wide range of difficulties is used to assign subjects to one of several measurement tests with item difficulties peaked around specific points along the hypothesized ability continuum.

Figure 3 (from Bayroff, 1964) depicts an 8-item routing test coupled with a 6-item measurement test.

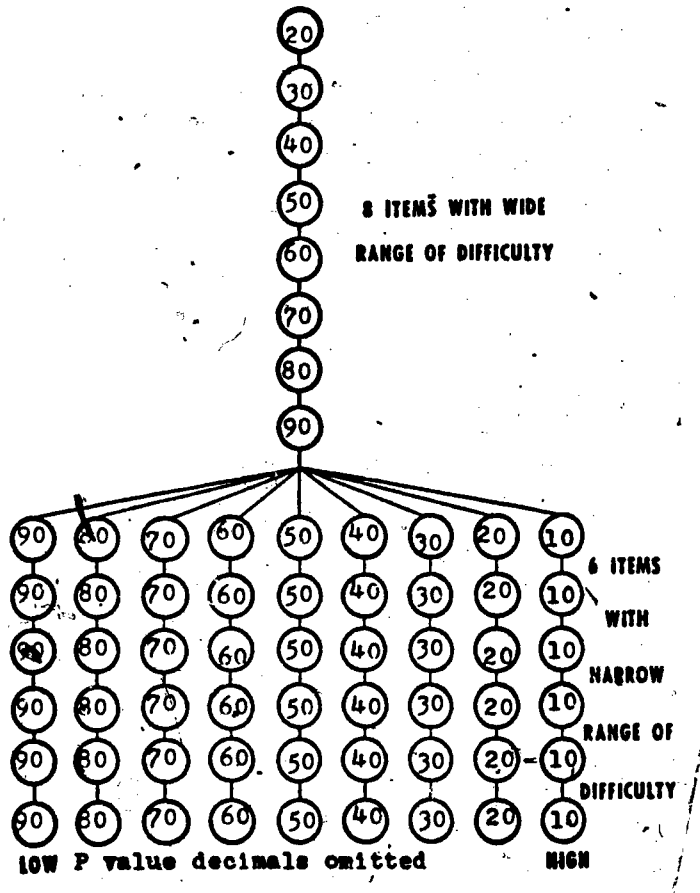


Figure 3. **Bayroff's example two stage adaptive test (1964).**

A subject's score was determined as a direct function of the number of correct responses or as a function of the item difficulty and discrimination of those items answered correctly.

In the majority of multi-stage adaptive testing research, a pyramidal model similar to that depicted in Figure 4 has been followed. In the example shown in Figure 4, an 8-stage strategy was utilized. All subjects received 8 items, beginning with Item 1, which was generally the item of median difficulty. The change in item difficulty between stages (step size) was fixed (.05 in the example).

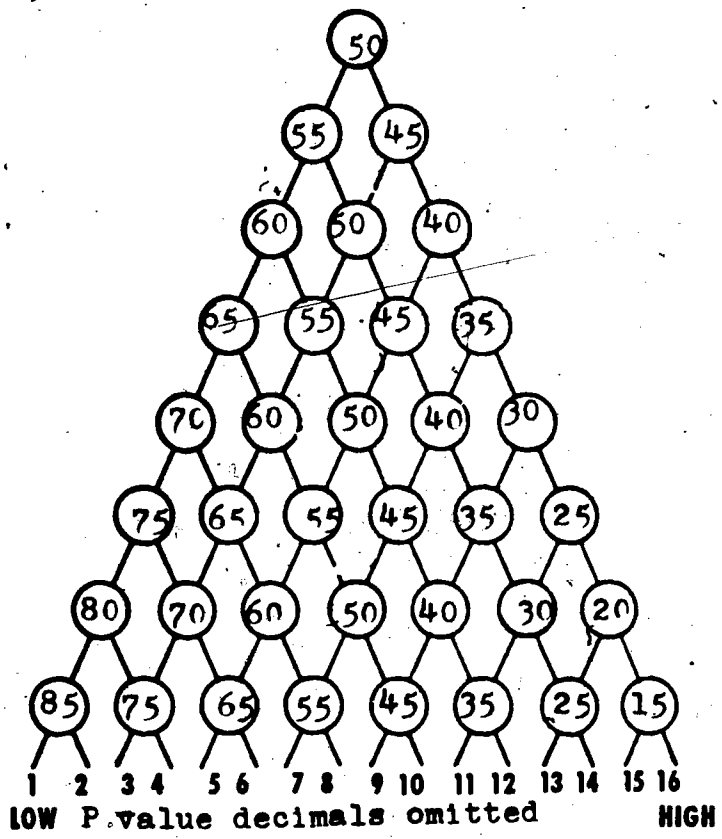


Figure 4. **Example of 8-step pyramidal adaptive test. (From Bayroff, 1964).**

A subject's score was based upon either the average difficulty of items answered correctly or upon the final item in the pyramid as shown in Figure 4. In this example a score ranging from 1 to 16 was assigned to the examinee.

The eight empirical studies in this cell of Table 2 reached general concensus in research results. All but Olivier (1974) and Wood (1969) found increases in the precision of measurement utilizing adaptive testing. Olivier attributed his result to unaccounted variance in the test scores possibly being caused by unfamiliarity of the subjects to the flexilevel testing format. Wood's research utilized a paper and pencil branching technique which, like the flexilevel procedure, likely led to a large number of subjects branching incorrectly.

Of the eight studies in this cell, the correlation between the short adaptive test scores and the longer conventional scores were in the .78 to .86 range with the exception of Wood's pooled results showing only a .51 relationship. As a group, these studies tended to recommend further research in adaptive testing be centered in mechanical or computer-based modes of presentation rather than the traditional paper and pencil method. The five papers utilizing such equipment all suggested further research in the area of adaptive testing.

Lord presents a discussion of tailored testing theory in general in Holtzman (1970). He provided a brief description of item characteristic curve theory, information function theory, several strategies of step size variation, several suggested scoring methods for tailored testing and varied number of items. He included in the final section of the paper the following caveat:

> If, for example, 500 items are available for tailored testing, better measurement will often be obtained by selecting, for example, the N=60 most discriminating items (highest a_i) and administering these as a conventional test, rather than using all 500 in a tailored testing procedure. *This may actually prove to be a fatal objection to any general use of tailored testing.* (Emphasis Lord's).

It is the judgment of the author of this paper that the Lord (1970) paper should be essential reading for any researcher interested in adaptive testing. Although the majority of adaptive testing research reported to date appears promising, Lord's warning should be kept in mind when evaluating the effectiveness of any adaptive testing strategy.

Variable Step Sizes

Theoretical studies. The majority of the theoretical research into a fixed number of stages and variable step sizes has been under the Robbins-Munro branching rule. Stocking (1969) and Lord (1970, 1971c, 1971d) have analyzed the Robbins-Munro technique in comparison with the more conventional up-and-down method described by Lord (1970). Essentially the Robbins-Munro, or so-called shrinking step size method, presents an item of median difficulty (b_1) to begin the test. If item b_1 is correctly answered, item b_2 is selected thusly:

$$b_{i+1} = b_i + d_i (u_i - \partial) \qquad (1)$$

(From Lord, 1971c)

where: $d_1, d_2, d_3, \ldots$ is a decreasing sequence of positive numbers chosen in advance of testing.

b_1 = difficulty of the i^{th} item.

$u_i = 1$, if item i is answered correctly and

$u_i = 0$, otherwise

∂ and d are positive numbers to be chosen prior to testing in order to produce good measurement properties on the final test scores.

The fixed step size methods discussed earlier determine the difficulty of the (i + 1)th item by a constant increment independent of i:

$$b_{i+1} = b_i + 2d (u_i - \partial) \qquad (2)$$

Lord (1970, 1971c, 1971d) compared these two step size strategies and found that the shrinking-step sizes provided better measurement than several varieties of up-and-down methods. A major deterrent to the shrinking-step sizes was reported. The up-and-down method requires an item pool of only n (n + 2)/2 items (for a 15 item test, for example, a 120-item pool is necessary) which is reasonable in most large scale testing situations. To use the Robbins-Munro strategy, $2^n - 1$ items should be available (32,767 items for a 15 item test) literally an impossibility in any situation. Since both empirical and theoretical research (Wood, 1971; Novick, 1969) have shown with a remarkable degree of consistency that adaptive testing is most effective between 15 and 20 items per test, the shrinking-step size methods as now conceptualized are not feasible in the

real world despite their theoretical superiority. In reality, Lord found this superiority to be relatively small and recommended use of the fixed step-size procedures rather than a Robbins-Munro whenever the number of items exceed six.

Lord (1970) and Stocking (1969) also investigated the persistent problem of how to score adaptive tests. Since different subjects may take different collections of test items in different orders, the conventional practice of rights-only or rights-corrected-for-guessing is clearly inappropriate. Lord's theoretical research showed that scores based upon the average difficulty of items answered correctly was superior to scoring methods based upon the difficulty of the final item passed or of the next item that the examinee would have taken. Conceptually, the latter two methods appear sound, since the estimate of the subject's true ability should improve as more items near the subject's .50 probability level are presented. If the subject's true score is far from b_1, the author would expect the early items faced by the subject to adversely affect average difficulty scoring methods. Certainly this area of adaptive testing remains to be empirically evaluated beyond Lord and Stocking's hypothetical investigations.

Simulated studies. Paterson's (1962) monte carlo study evolved from the sequential item test (SIT) of Krathwohl and Huyser (1956). A six-item conventional test and six-item pyramidal test were created with 1500 "examinee" scores generated at 15 different ability levels (100 each level). Unlike all of the other studies of adaptive testing reviewed, Paterson selected items based upon biserial correlation rather than exclusively by item difficulty. He ordered the items in the pool by difficulty and by r_{bis} within difficulty levels. Step size was thus a function of item discrimination, approximating a shrinking-step size model since larger steps were taken for early items and shorter step sizes for later items. He scored his tests based upon the final difficulty method. His results showed the adaptive test to better reflect non-normal ability distributions and to better measure examinees with abilities in the extremes of the distribution. As with Lord, Paterson found measurement efficiency slightly inferior for the adaptive strategy near the mean of the score distribution. He recommended that adaptive item pools required a more flat distribution of item difficulties than the conventional test.

The only other simulated study of fixed number of stages with variable step size adaptive testing was done by Bryson (1972). She compared two 5- and 10-item adaptive measures with two 5-item conventional tests with a validity coefficient based upon a 100 item parent test serving as criterion. Her results did not favor the adaptive procedure; however, several methodological errors involving branching, scoring and the fact that the control group tested via paper and pencil while the experimental group used a cathode ray tube (CRT), suggest the discounting of her results.

Bryson further compared her empirical results described above, with two groups of test scores of 100 recruits which were rescored *as if* they had been taken sequentially as Cleary, Linn, and Rock (1968a, 1968b) had done earlier. The correlation of these four group scores to the parent test yielded one group with higher adaptive correlation and one with higher conventional correlation. Such a result leads one to question the procedure of using "real data" from data banks for simulations of adaptive test results. Apparently, an interaction effect exists between item order, item selection and/or examinee response which invalidates this type of simulation design.

Empirical studies. The aforementioned paper of Bryson (1971) and two studies by Bayroff's associates (Seeley, Morton, and Anderson, 1962; Bayroff, Thomas, & Anderson, 1960) comprise the only reported empirical studies of adaptive testing with variable step sizes. The Bayroff studies incorporated one unique twist in adaptive research in that branching from the first item was based upon not only whether the subject's response was correct or not, but also upon the incorrect responses. The attempt at utilizing the "partial knowledge" information available to discriminate between examinee ability levels has been extensively investigated (review by Stanley & Wang, 1970) on an entire test basis with increases in test reliabilities and decreases in test validation generally reported. The major problem appears to be finding enough "good" items, all with "good" distractors to comprise a test. Under the Bayroff group's strategy, only one or two such items would be required, which seems much more feasible. Such an approach seems to be worthwhile for further investigation.

Results from the Bayroff studies showed a .63 correlation for a 6-item adaptive test with a parent test while a 25-item conventional test correlated significantly higher with a parent test. The authors noted that the distribution of item difficulties was badly skewed to the left with a resultant skewed score distribution. In addition, the adaptive tests involved longer construction, administration and scoring time and resulted in more unusuable answer sheets than the conventional tests. These results are consistent with the Wood (1969) and Olivier (1974) results using paper and pencil adaptive tests. Apparently, a mechanized mode or presentation should be used for any adaptive testing to avoid examinee branching errors.

Variable Number of Stages

Research studies on adaptive testing involving variable numbers of stages fall under the category of decision theory. In these studies, testing was terminated when a preset criterion was reached. Commencing with the work of Wald (1946) and the Statistical Research Group (SRG) and carried on by Cronbach and Gleser (1965), sequential analysis techniques entailed presenting an item or block of items to a subject, after which a decision is made to (a) assign the subject to a "passing" group; (b) assign the subject to a "failing" group; or (c) continue testing. All of the research done within the variable number of stages level essentially follow the sequential analysis model in determining a stopping rule for testing.

Conceptually, varying the number of items presented between subjects makes sense. Setting a particular number ot items for all subjects fails to account for individual differences between subjects and certainly must be wasteful for a percentage of the examinees. The catch, of course, is in determining when to cease testing for each subject and handling the problems which arise when examinees do not take an equal number of items.

Constant Step Size

Theoretical studies. About a decade after the previously cited work of Wald and the SRG, Cronbach and Gleser's (1965) book, *Psychological Tests and Personnel Decisions,* presented a complete theoretical exposition of efficient testing procedures. They introduced the concept of cost effectiveness and concluded that, theoretically, testing efficiency will be maximized by completely adapting the test to the individual testee. Green (1970) reiterated the cost effective point in responding to Lord's (1970) caveat concerning adaptive testing. Kappauf (1969) described an application of the up-and-down method of branching using a sequential analytic stopping rule for computer-based psychological testing, although no results were reported. No further theoretical research was found until Weiss (1973) presented his model he termed "stradaptive testing," produced under a research grant from the U.S. Navy to investigate computer-based adaptive testing for possible Navy implementation on a large scale. Weiss and his associates are in the process of comparing two-stage, Bayesian, pyramidal, flexilevel and stradaptive testing strategies with one another and with conventional testing. DeWitt and Weiss (1974) published a description of an elaborate computer software system for making these comparisons and McBride and Weiss (1974) produced a description of the mechanics of creating an item pool for adaptive testing research. Since Weiss' stradaptive model is the target of this present study, the description of the model will be held until Section III of this paper when a complete definition of the elements of the model will be made.

Simulated studies. The author found only one simulated study involving constant step sizes and a variable number of stages. Linn, Rock, and Cleary (1970) reanalyzed 1967 College Level Examination Program (CLEP) data from English composition, mathematics and natural sciences examinations. They simulated two adaptive testing strategies, one in which three CLEP tests were analyzed separately and the other in which the mathematics test score was used in the decision process for the English and science tests. Essentially, Linn *et al.*, followed the sequential analytic procedures suggested by Wald, although the specific model was developed by Armitage (1950). They also scored short conventional tests of the first 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, and 60 items for comparisons with the adaptive tests.

Linn, *et al.*, results showed substantial improvement in assignment of subjects to one of two groups for dichotomous decision making. They estimated that the short conventional tests required approximately twice as many items to achieve a comparable level of accuracy as that achieved by the adaptive tests. To the author's knowledge, no empirical study has been conducted to verify this impressive result. Such a study is warranted, since other "real data" simulation results have not replicated empirically.

Empirical studies. No published research on adaptive testing with constant step sizes and a variable number of stages was found by the author with the exception of examples of stradaptive records reported in Weiss (1973). Weiss is presently investigating this area and has advised the present author (personal communication) of some aspects of his results. Weiss' test-retest reliabilities on ten different scoring methods have been in the range of .72 to .93 for a method which branches the subject to an easier item whenever he either misses the previous item or responds with a question mark. Weiss' alternate stradaptive testing model (which is the model used in the present study) presents another item of equal difficulty after a question mark is entered in response to an item. His resulting test-retest reliabilities using this model have been consistently about .10 lower than that from the other model.

Two empirical studies have been made (Cowden, 1946; Moonan, 1950) which verified the sequential analysis application in testing. However, the tests used were presented to the subjects in a fixed order, with only the number of items being presented being varied. This strategy is not adaptive testing, per se. Thus, these two studies have not been included in Table 1. The favorable results do provide evidence that an increase in testing efficiency is possible by adapting the number of items on a test to the individual subject.

Variable Step Sizes

Models in which both the number of stages and step sizes are variable generally fall under the heading of Bayesian testing. All reported work in the area has been published during the last five years. Computer implementation seems essential since the selection of each item for a given examinee takes into account all previous responses. A criterion is established such as to minimize measurement error by providing an estimate of the subject's ability. This estimate is a weighted average of the norming group's performance on an item and the subject's performance on the items taken up to that item.

Theoretical studies. Two models have been suggested for implementing the Bayesian formulas in adaptive testing. Novick (1969) and Owen (1969, 1970) have produced radically different models which appear to be conceptually appealing. The complexity of the Bayesian models prohibits lengthy description in this paper. However, some of the results have direct application to more conventional adaptive testing. Novick (1969) anticipated Bayesian testing to be particularly advantageous for tests of 15 to 20 items of length. This result has been supported in the fixed number of stages empirical studies reviewed earlier and also in Wood's (1971) empirical study of Owen's model. This consistency of results in the adaptive testing literature provides strong evidence of the potential savings in the number of items required in adaptive testing.

Simulated studies. Urry (1970, 1971) has reported two monte carlo studies of a model based upon a logistic test model. Like the Bayesian models, Urry's strategy chose items in order to minimize the standard error of the estimate of the subject's ability. Unlike Bayesian testing however, Urry's model utilizes maximum likelihood estimates calculated after each item to estimate ability.

Urry varied item-ability biserial correlations, number of items, difficulties, the guessing parameter and the shape of the distribution of item difficulties to generate 36 item structures. His criterion was validity in the prediction of the scores of 100 hypothetical "subjects" of known ability levels.

His results showed his adaptive tests to be increasingly effective when item discrimination increased, particularly when a broad range of item difficulties was used. In such a situation, he found a 10-item adaptive test to be as effective as a 30-item conventional test. He also suggested that adaptive testing not be used when the probability of guessing an item correctly is .50, as in a true-false test.

Urry's results also indicated that when high item discrimination indices were coupled with a rectangular distribution of difficulties, a 10-item adaptive test produced as high a correlation between known and estimated ability as a 100-item peaked test. When he analyzed the results with item discrimination set at .45 such as Lord (1970, 1971) used, his results confirmed Lord's less dramatic conclusions. He concluded that adaptive testing be used when item ability biserial correlations are .65 or larger. Unfortunately, a large pool of items above this criterion would be most unusual in the typical ability testing situation. If such a minimum standard is necessary for adaptive testing to be empirically effective, this fact alone could toll the death knell for this testing strategy.

Urry's second study (1971) used the same model as his earlier dissertation. He generated three item banks and fit the data to the model. He determined that Bayesian testing of the Verbal Scholastic Aptitude Test (VSAT) could save 65% of testing time for the average examinee.

Kalisch (1974) used the beta distribution and conditional item difficulties to predict subject responses on items beyond those he actually took. A sequential decision rule was used to determine when to cease testing based on an expected loss function to the subject between the three possible decisions (item response would be correct, no assumption, or response would be incorrect). Results were reported as favorable to future research into this model.

Empirical studies. Four empirical studies of fully adaptive testing have been reported. Wood (1969) conducted an empirical validation (number of subjects only 28) of Owens' (1970) model along with a simulated study as part of a dissertation. In the simulation portion of the study, he compared his Bayesian results with a 60-item simulated two-stage test and a 60-item conventional test. The empirical data showed the Bayesian ability estimates to converge around 20 items, remarkably similar to Novick's (1969) theoretical prediction with a different Bayesian testing model. In the simulation portion of the study, Wood found both Bayesian and two-stage testing to be superior to conventional testing, with the two-stage performing better than the Owen model in terms of measurement preciseness, although the Bayesian method was more cost effective. A saving of 2/3 of the number of items required for the conventional and two-stage tests was evidenced in the results of the Bayesian strategy. This result also supported Owens' theoretical savings.

Ferguson's dissertation (1969) and a later paper (1971) report a model development and empirical validation for a computer-assisted, criterion-referenced instructional system. The purpose of his research was to apply the sequential analytic techniques of Wald to the decision of mastery or non-mastery of instructional objectives within a hierarchially-structured domain of achievement. After each item response a decision was made to classify the student as having mastered the material, not mastered it, or no decision (present another item). Testing continued until a decision was reached for all students. The computer then selected the next objective for each subject based upon previous performance.

Ferguson's results were very favorable to the adaptive approach. Both test-retest reliabilities and validities were higher than a conventional paper and pencil mode of presentation and a 50% time savings was reported on the computer-based measurement system.

As Green suggested (1970) and Ferguson's research confirmed, the use of adaptive testing as a strategy for instructional management rather than as a measurement tool may turn out to be the most effective application of the adaptive models. The instructional situation is immediately concerned with decisions about a single subject and the oft-mentioned lack of efficiency of the adaptive strategies near the center of the ability distribution should not be entirely relevant in this context. Further research into instructional applications of adaptive testing is warranted.

Summary of the Literature on Adaptive Testing

The following conclusions appear warranted based upon the studies in this review:

1. Item pool distributions of difficulty and discrimination values have a large effect on empirical results in adaptive testing studies. Well-normed item statistics with appropriate distributions are essential for empirical studies.

2. Average difficulty scoring methods are superior to final difficulty methods.

3. Within the fixed number of stages dimension, the up-and-down method is superior to the Robbins-Munro method due to the number of items required in the item pool.

4. At least with the models developed to date, paper-and-pencil adaptive testing is not likely to produce favorable results. Use of a computer greatly enhances this measurement strategy.

5. Although an efficient method for analyzing a model, "real data" simulation studies should be followed up by empirical validation. The change of item sequencing, item content and test length in adaptive

testing apparently affects examinee performance. This change, at least in the studies reviewed, was consistent—the simulated studies were far more favorable to adaptive testing than the empirical validations of the same model.

6. Theoretical studies need to consider item parameters more closely attuned to the reality of measurement. Although assumptions of no guessing, all items having equal difficulty or discrimination indices, etc., simplify analysis, the results of this type study are not generalizable to the world of testing. Follow-up validations are essential.

7. Group indices such as reliability and validity may not be appropriate measures of the effectiveness of adaptive testing. An information function as described by Lord seems preferable.

8. A fully adaptive model in which both the number of items presented and a variable step size should produce the greatest gains.

9. A large reduction in the number of items necessary for effective measurement seems probable using adaptive procedures.

10. Adaptive testing shows promise as an effective, feasible alternative to conventional testing.

III. THE STRADAPTIVE TESTING MODEL

Lord's theoretical analysis of adaptive testing versus conventional testing made one point very clear ... a peaked test always provided more precise measurement than an adaptive test of the same length *when the testee's ability was at the point at which the conventional test was peaked.* As shown in Figure 2, at some point on the ability continuum, generally beyond about + .5 standard deviations from the mean, the adaptive test requires less items for comparable measurement efficiency.

Lord's conclusion suggests that an "ideal" testing strategy would present a collection of items to each subject comprising a peaked test with a .50 probability of a correct answer for examinees of the particular subject's true ability (P_c = .50). The catch, of course, is that the true ability of the subject is unknown; the estimation of which is, in fact, the desired outcome of the measurement procedure.

Traditionally, this problem has been circumvented by peaking the test at P_c = .50 for the hypothetical average ability level subject. This procedure worked well for examinees near the center of the ability continuum, but less efficiently near the extremes.

Weiss and colleagues at the University of Minnesota have developed and begun validating a model designed to combine the best of both of these two competing measurement strategies. They have combined the underlying philosophy of the Binet-Simon IQ measurement with the work of Lord to produce their so-called stradaptive testing model (stratified adaptive). The Binet testing procedure began testing at an "entry point" on the ability continuum judged to be appropriate by the examiner. He presented a short sub-test to the subject which was peaked around P_c = .50 for subjects of a comparable "mental age." Based upon the subject's proportion of correct responses to the first sub-test the examiner selected the next peaked sub-test which had an average P_c = .50 for groups of respectively higher or lower mental ages.

The Binet strategy defined two subtest levels for a subject. [illegible]he early testing, the examiner searched for the subject's "basal age," that is, the peaked test in which the examinee answered all items correctly. Determination of an examinee's basal age assumed that any less difficult peaked tests would also be below the subject's true ability level, thus providing a lower bound on the true ability estimate. Once the basal age is found, the Binet examiner selects progressively more difficult subtests until the subject's "ceiling age" is defined. The ceiling age was determined by the subtest in which the subject incorrectly responded to all items. Testing beyond this difficulty level would only frustrate the subject, reducing the precision of measurement. It was assumed that any item more difficult than the subject's ceiling level would similarly have been answered incorrectly. The items between the basal and ceiling ages provided accurate ability estimation for the subject. If the subtests had been properly normed, the subject's proportion of correct responses within the subtests he had taken should decrease monotonically from 1.00 at his basal age to 0.00 at his ceiling age. The best estimate of his true ability would be a function of the difficulty of that subtest in which his $P_c \cong .50$.

Weiss' stradaptive model extends this Binet rationale to computer-based ability measurement. A large item pool is used with the item parameter estimates based on a large sample of subjects from the same

population as the intended examinees. The items are scaled into a set of peaked levels (strata) according to their difficulties. The subject's first item is selected based upon a previously collected ability estimate or the subject's own estimation of his ability on the dimension being assessed.

As in the Binet, the subject's basal and ceiling strata are defined, with testing ceasing when the ceiling stratum is determined. A subject's score is a function of the difficulty of the items answered correctly.

The Item Bank

A stratified, assumed unidimensional, item pool is required for a stradaptive test. Items are organized into a number of strata peaked at different difficulty levels.

Weiss (1973) lists four steps in the creation of the item pool for a stradaptive test.

1. Test a large number of subjects on a large number of items which measures an hypothesized unidimensional trait.

2. Compute item difficulty and discrimination indices on all items in the item bank, in either traditional p-values and item/total score correlations or using latent trait theory parameter estimates derived from normal ogive item assumptions (Lord & Novick, 1968). The latter alternative is preferable if the assumptions of the normal ogive model can be accepted since, theoretically, the estimates derived from this model are not contingent upon the frequency distribution of ability of the total group. That is, the item characteristic function is the same for any group of examinees on the unidimensional trait of concern. Two assumptions underlie latent trait theory: 1) the latent variable space is one-dimensional (K = 1) and 2) the metric for the ability continuum (θ) can be chosen so that the item characteristic curve for each item g = 1,2, . . .,n (the regression of item score on θ) is the normal ogive

$$P_g(\theta) \equiv P_g(\theta, a_g, b_g) \equiv \Phi(L_g(\theta)) \equiv \int_{-\infty}^{L_g(\theta)} \varsigma(t)\, dt = \int_{L_g(\theta)}^{\infty} \varsigma(t) dt,$$

where

$$L_g(\theta) \equiv a_g(\theta - b_g)$$

is a linear function of θ involving two item parameters a_g and b_g, and $\varsigma(t)$ is the normal frequency function. See Lord and Novick (1968) chapter 16 for further discussion of the normal ogive model and latent trait theory.

3. Assign the items in the pool into I independent strata, where each stratum is a peaked test of J items with no overlap of item difficulties between adjoining strata. The number of strata, I, depends on the size and distribution of item difficulties, with the precision of measurement approaching equality throughout the distribution of ability levels as I increases. Figure 5 depicts the item pool stratification plan.

Weiss recommended that a minimum of 10 to 15 items per stratum appeared appropriate and that experience with the model suggested more items be placed in the lower and middle difficulty strata than at the upper strata.

4. Arrange the items within strata by discrimination index from top to bottom in each stratum. Since items taken earlier in a stratum should reflect a wider range of abilities, finer discrimination is not required. Items lower in a stratum should be reached when testing is confined to only a narrow range of abilities and "fine" discrimination between ability estimates is necessary.

Table 3 shows the actual distribution of items used in this experiment. The final pool included 244 items grouped into 9 strata according to normal ogive item difficulty parameters as shown in Table 3.

Figure 6 shows the relationship between A_g and B_g parameters in the stradaptive pool. As is typical in educational and psychological research, the concentration of more difficult items contain the lower discrimination values. The correlation between b_g and a_g of –.31 reflects this problem. Selection and rescaling procedures will be described in Chapter four of this paper.

The nine strata in Table 3 are essentially nine peaked tests varying in average difficulty from – 2.12 to + 1.91. Stratum 9, the most difficult peaked test, for example, was composed of 19 items ranging from b_g = 1.27 to b_g = 3.68. The order of items within a stratum was random, unlike Weiss' model, in order to permit an

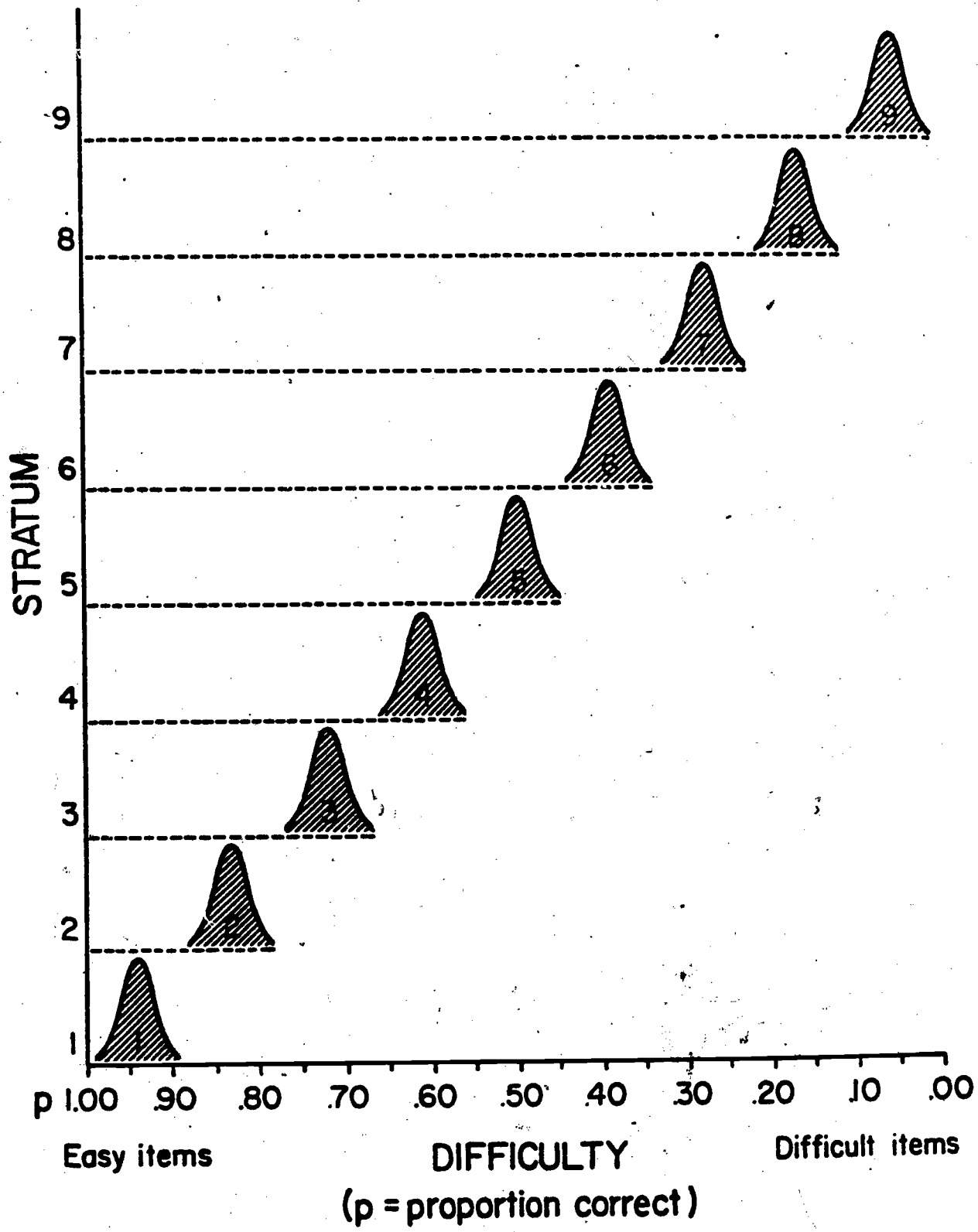


Figure 5. **Distribution of items, by difficulty level, in a stradaptive test (from Weiss, 1973).**

Table 3

Item difficulties (b) and discriminations (a), based on normal ogive parameter estimates, for the Stradaptive Test item pool

	Stratum (easy)								(difficult)
	1	2	3	4	5	6	7	8	9
Item Difficulties									
High	-1.94	-1.46	- .90	- .49	- .10	.25	.67	1.34	3.68
Low	-3.57	-1.91	-1.40	- .88	- .44	- .10	.27	.71	1.27
Mean	-2.12	-1.68	-1.13	- .68	- .25	.04	.44	.95	1.91
No. of Items	20	26	33	39	31	28	26	22	19

Item Number Within Stratum	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b	a
1	-2.08	.48	-1.87	.50	- .90	.56	- .62	.79	- .19	.29	.20	.59	.38	.53	1.25	.46	1.76	.49
2	-1.97	.45	-1.74	.69	-1.05	.83	- .49	.69	- .33	.41	.25	.59	.31	.34	.76	.64	1.69	.44
3	-2.07	.42	-1.70	.95	-1.34	.42	- .72	.77	- .11	.50	.00	.44	.43	.53	1.19	.56	1.61	.44
4	-2.27	.64	-1.91	.52	-1.11	1.17	- .65	.73	- .17	.50	.24	.34	.63	.39	.81	.45	2.91	.49
5	-1.97	.86	-1.50	.77	-1.39	.63	- .88	.49	- .11	.53	.09	.69	.65	.66	1.13	.36	3.69	.28
6	-2.17	.36	-1.79	.59	- .92	.50	- .49	.61	- .16	.33	- .10	.61	.34	.49	.87	.34	1.57	.29
7	-2.31	.41	-1.47	.86	-1.06	.42	- .80	.50	- .16	.83	.09	.71	.30	.71	.71	.48	1.60	.33
8	-2.03	.41	-1.83	.55	-1.31	.44	- .69	.75	- .11	.52	.12	.81	.59	.53	.88	.53	1.34	.42
9	-2.13	.48	-1.68	.58	-1.22	.95	- .55	.52	- .44	.69	.11	.52	.28	.64	.88	.49	1.83	.52
10	-3.57	.30	-1.52	.93	-1.08	.52	- .80	.55	- .42	.55	.00	.39	.38	.52	.79	.61	1.27	.77
11	-2.03	.50	-1.69	.83	-1.19	.79	- .57	.33	- .21	.39	.00	.41	.29	.61	1.24	.30	2.29	.44
12	-2.63	.39	-1.69	.41	- .95	1.01	- .84	.69	- .35	.59	.21	.53	.62	.56	.71	.71	1.33	.33
13	-1.95	.25	-1.65	.71	-1.37	.53	- .86	.83	- .24	.79	.13	.68	.53	.55	.91	.26	1.91	.40
14	-1.95	.56	-1.56	.64	-1.31	.64	- .76	.59	- .10	.55	- .08	.77	.29	.37	1.06	.49	1.27	.42
15	-2.31	.63	-1.90	.69	-1.40	.75	- .54	.46	- .42	.75	.13	.44	.27	.68	1.24	.33	1.91	.27
16	-2.50	.53	-1.51	.88	- .90	.36	- .53	.73	- .41	.66	.00	.71	.56	.45	1.01	.56	2.94	.25
17	-2.03	.50	-1.88	.59	-1.04	.68	- .83	.58	- .16	.83	- .05	.56	.67	.46	.75	.79	1.94	.41
18	-2.36	.61	-1.83	.90	- .97	.81	- .51	.58	- .30	.58	.13	.44	.40	.59	1.34	.37	2.13	.27
19	-1.95	.81	-1.80	.36	-1.09	.68	- .62	.79	- .31	.34	.14	.66	.32	.53	.95	.25	1.33	.37
20	-2.03	.71	-1.55	.61	- .91	.77	- .86	.55	- .31	.45	.05	.64	.30	.73	.75	.66		
21			-1.65	.45	-1.02	.75	- .64	.68	- .18	.68	- .91*	.97	.29	.48	.79	.46		
22			-1.78	.68	-1.18	.46	- .85	.46	- .33	.64	- .06	.50	.66	.64	.94	.53		
23			-1.50	.77	-1.35	.45	- .59	.77	- .35	.69	.12	.77	.37	.66				
24			-1.46	.63	-1.17	.58	- .53	.41	- .18	.48	.06	.50	.56	.70				
25			-1.46	.49	-1.07	.27	- .65	.66	- .44	.52	.10	.55	.50	.68				
26			-1.90	.79	- .95	.66	- .75	.73	- .16	.81	.00	.45	.56	.39				
27					-1.36	.98	- .54	.88	- .23	.49	- .04	.88						
28					-1.27	.71	- .60	1.07	- .19	.44	.07	.36						
29					-1.39	.88	- .74	.61	- .37	.79								
30					- .90	.71	- .61	.64	- .14	.66								
31					-1.30	.69	- .83	.81	- .18	.48								
32					-1.38	.36	- .75	.73										
33					-1.21	.45	- .60	.59										
34							- .88	.81										
35							- .77	.48										
36							- .49	.33										
37							- .65	.33										
38							- .76	.40										
39							- .73	.83										

*This item was misassigned to stratum 6 rather than 3. Fortunately, no subjects reached the item in the Stradaptive Pool.

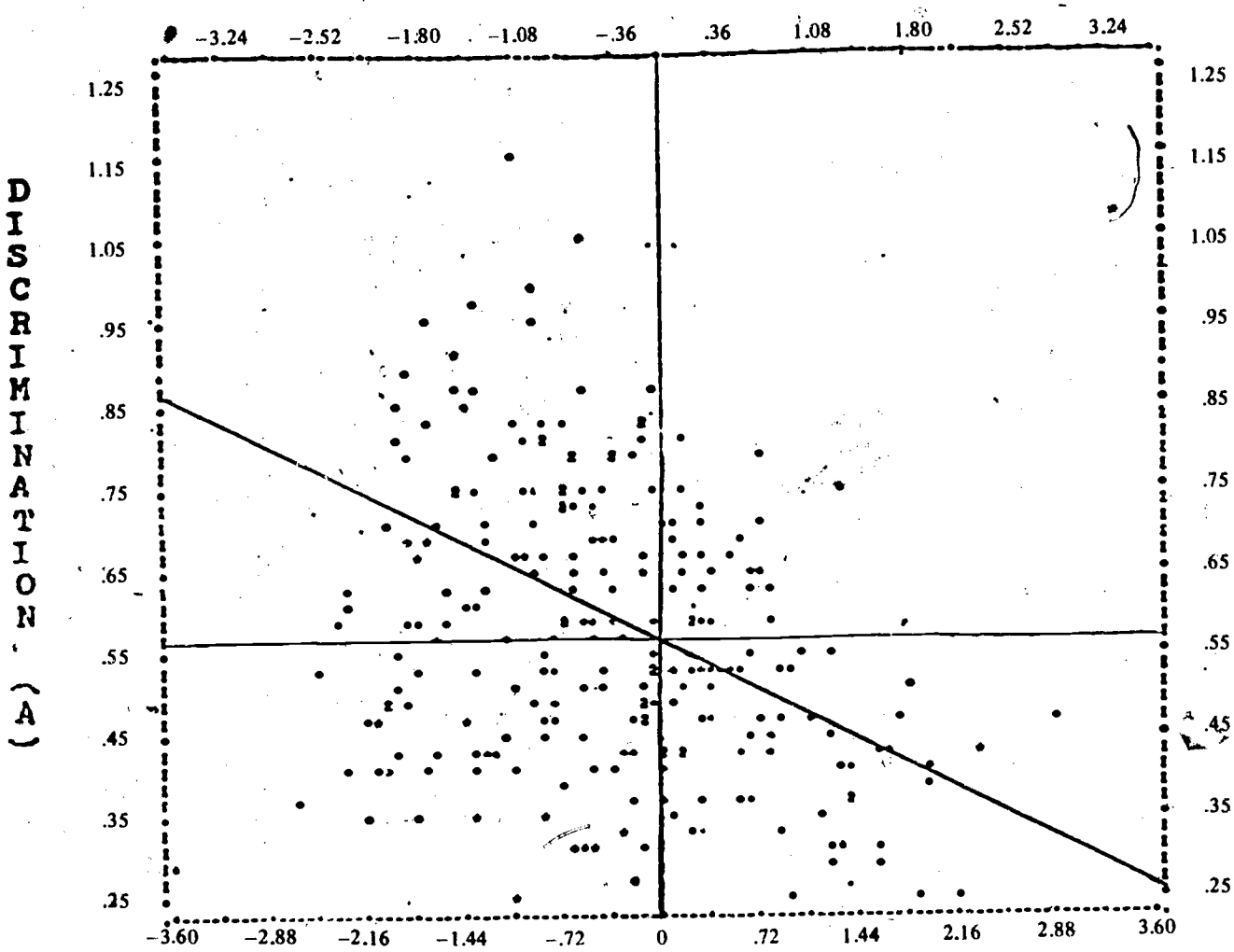


Figure 6. **Scatterplot of relationship between A_g and B_g.**

alternate-forms reliability coefficient to be calculated on stradaptive examinees. Personal discussion with Weiss led to the conclusion that the randomized design utilized in this study would not jeopardize the feasibility of the stradaptive testing procedure. Theoretically, this design could have added a few items to some examinees' tests, although ability estimates should have been similar to Weiss' procedure estimates. If a bias were caused by this change, it would make the results from this study *less* impressive than they might be otherwise in a comparison between stradaptive and conventional testing.

Item Content and Format

All items in the item pool were in the following form:

EXAMPLE:	Calf:	Cow:
	a. puppy:	dog
	b. nest:	bird
	c. house:	build
	d. shell:	turtle

These test items were selected for this study for a number of reasons. First, the SCAT Series II provided a single-format unidimensional test with extensively-normed item parameter estimates. The item format was

easily stored in the computer item file, being short and standard for all 244 items. SCAT II was well received in Buros' *7th Mental Measurements Yearbook* (1971) with internal consistency reliabilities for the five 50-item forms ranging from .86 to .88 and validities comparable to other leading measures of verbal aptitude. Administration was relatively short (20 minutes for the published test) and, finally, ETS consented to provide the items and item parameter estimates for this research.[1]

Computer Program for Model Implementation

A computer program fully described by DeWitt and Weiss (1974) was adapted by James Sutherland of Florida State University to fit the FSU Control Data Corporation 6500 computer.[2]

Instructional Sequence

The DeWitt and Weiss program was written so that it could be used by subjects with no prior cathode-ray-tube experience and with no help from the examination proctor. The proctor simply typed a single letter into the CRT to select stradaptive or conventional test, and the instructional sequence began. The subject was asked to type in his social security number and name and was instructed in the use of the CRT and in the nature of the research. A sample item was presented and responses to the questions in Figure 7 were requested.

Everybody is better at some things than others . . . Compared to other people your age, how good do you think your vocabulary is?		Entry Stratum (not seen by examinee)
Better than:	1 out of 10	 1
	2 out of 10	 2
	3 out of 10	 3
	4 out of 10	 4
	5 out of 10	 5
	6 out of 10	 6
	7 out of 10	 7
	8 out of 10	 8
	9 out of 10	 9
Type in the number from 1 to 9 that gives the number of people you would guess you are better than (in vocabulary).		

Figure 7. **Entry point question for determining subject ability estimate (from: Weiss, 1973).**

After completing this task, the subject typed in the word "start" and the testing sequence began.

Testing Sequence

The response to the question in Figure 7 determined the subject's entry point (ability estimate) in the stradaptive item matrix. The first item the stradaptive subject received was the first item in the stratum commensurate with his ability estimate. The subject was then branched to the first item in the next higher or lower stratum depending upon whether the initial response was correct or incorrect. If the subject entered a question mark (?), the next item in the same stratum was then presented.

---

[1]The test materials from the SCAT Series II Verbal Ability tests were adapted and used with the permission of Educational Testing Service. The author of this paper gratefully acknowledges the help of ETS in the pursuit of this research.

[2]DeWitt's help in the conversion of his program from the University of Minnesota system to the Florida State University system is gratefully acknowledged. Under the time constraints in this study, program operation prior to data collection would not have been possible without DeWitt's advice and efforts in our behalf.

Testing continued until a subject's ceiling stratum was identified. For this study, the ceiling stratum was defined as the lowest stratum in which 25% or less of the items attempted were answered correctly, with a constraint that at least 5 items be taken in the ceiling stratum. The 25% figure reflects the probability of getting an item right by random guessing on a 4-option multiple choice test. Once a subject's ceiling stratum was defined, the program looped back to the examinee's ability estimate stratum and commenced a second stradaptive test with item selection continuing down the item matrix from where the first test ended. Since items were randomly positioned within each stratum, parallel, alternate forms were taken by all subjects who reached termination criterion on the first test.

A maximum of 60 items per subject per test was established, as pre-study trial testing suggested that subjects became saturated beyond this point.

Scoring Methods

Weiss (1973) suggested ten possible scoring methods for stradaptive testing. These scoring methods equate item difficulties to ability estimates through the scaling to normal ogive parameters, assuming a unidimensional continuum underlying the item pool.

Most of Weiss' scoring method suggestions were used in this study unchanged. The item scoring methods can be classified into three types: item scores, stratum scores, and average difficulty scores.

Highest Item Difficulty Scores. Three scoring methods are based on the "hurdle" concept in ability measurement: that is, the height (difficulty) of the highest hurdle a subject can jump. Thus, a subject's ability can be estimated as:

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

Method 2. The difficulty of the n + 1th item (the next item that would have been presented if testing continued).

Method 3. The difficulty of the most difficult item answered correctly below the subject's ceiling stratum.

Stratum scores. Since the stradaptive pool can be considered a series of peaked tests, the average difficulty of the items within each of the strata is a measure of examinee ability for subjects whose ability lies within a strata. This rationale suggests four stratum scoring methods similar to methods 1 through 3. A subject's ability score can be estimated by:

Method 4. The average difficulty of the highest stratum in which at least one item was answered correctly.

Method 5. The average stratum difficulty of the n + 1th item.

Method 6. The average item difficulty at the stratum just below the ceiling stratum.

Scoring method 7 (the interpolated stratum difficulty score) weights method 6 by the P_c at the highest non-chance stratum, thus resulting in a continuous range of ability estimates.

Method 7. This scoring method is defined as:

$$A = D_{c-1} + S(P_{c-1} - .50)$$

where

1	D_{c-1}	=	the average difficulty of the c−1th stratum where c is the ceiling stratum
	P_{c-1}	=	the subject's proportion answered correctly at the c−1th stratum
and	S	=	$D_c - D_{c-1}$, if $P_{c-1} > .50$
or	S	=	$D_{c-1} - D_{c-2}$, if $P_{c-1} < .50$

where

D	=	average difficulty of the designated stratum.

This scoring method makes an assumption that the subject's ability lies at the mean difficulty of a peaked test (stratum) if exactly 50% of the items are answered correctly. Ability is estimated proportionally between the midpoint of his $C-1^{th}$ and C^{th} strata.

Unlike the other 3 stratum scoring methods, method 7 results in a hypothetical continuous range of possible scores along the entire continuum of ability.

Average difficulty scores. Three possible scoring methods are analogous to Lord's average difficulty methods. They estimate a subject's ability to be:

Method 8. The average difficulty of all of the correctly answered items.

Method 9. The average difficulty of all items answered correctly between the basal stratum and the ceiling stratum.

The scoring of method 9 was redefined in this study from Weiss' original definition. As specified by Weiss, method 9 was not usable when basal and ceiling strata were adjoining. When this result occurred in the present study, score 9 was defined as:

$$A = D_b + S(P_b)$$

where D_b = average difficulty of items answered correctly in basal stratum

and $S = D_c - D_{c-1}$

Method 10. The average difficulty of items correctly answered in the highest non-chance stratum.

Two other revisions were made by the author to Weiss' scoring suggestions. If no basal ceiling was established (i.e., no stratum emerged with 100% correct responses), it was assumed that the subject's basal stratum lay immediately below the lowest stratum with a correct response in it. Similarly, if no ceiling stratum was defined (i.e., the subject scored above 25% correct in all strata utilized), the subject's ceiling strata was assumed to be immediately above the highest non-chance stratum.

The author made one other change in the Weiss model. Weiss had reported (1973) a problem wherein subjects of extremely high ability "topped out" his test and answered a high percentage of the presented items in stratum 9 correctly. Hence, an amendment to the 5 item/25% termination criterion was needed.

Since the probability of a subject of true ability less than the average difficulty of stratum 9 correctly answering a stratum 9 item is <.50, the joint probability of such an individual correctly answering 5 items in stratum 9 *in a row* is < .05, the alpha level used throughout this research. Therefore, whenever 5 items in a row were correct in stratum 9, testing was terminated. The subject's basal stratum was not affected by the earlier termination, but his ceiling stratum became "stratum 10," whose mean difficulty was:

$$D_{10} = D_9 + (D_9 - D_8)$$

where D_i = mean difficulty of all items in stratum i

This change resulted in ability estimates for examinees in this category theoretically ranging from 2.27 to 3.75 for scoring methods 9 and 10. Such ability estimates would seem to be appropriate for subjects demonstrating such a strong response pattern.

Termination Rules

As indicated earlier, Weiss had two versions of his stradaptive testing computer program. Version one, which was used in this study, presented another item in the same stratum when a subject skipped an item.

The author of this study was unaware of the existence of the second branching strategy program prior to completion of data collection. However Weiss' program procedure of ignoring skipped items in determining test termination was questioned. It appeared that valuable information was being lost when the Weiss procedure was followed.

It was reasonable to expect that a subject would omit an item *only* which he felt he had no real knowledge of the correct answer. Thus, investigation of the ten scoring methods with termination based upon omits counted as wrong answers was judged appropriate.

Weiss had set 5 items in the ceiling stratum as the minimum constraint upon termination. A secondary goal of the present study was to determine what effect the reduction of this constraint to 4 would have upon the effectiveness of the 10 scoring methods in the stradaptive tests.

These two questions of the handling of omits and the variation in the constraint on the termination of testing created the following three methods for comparisons:

Termination Method 1: Omits ignored/constraint = 5 items

Termination Method 2: Omits = wrong/constraint = 5 items

Termination Method 3: Omits = wrong/constraint = 4 items

Data was collected using termination Method 1 and then rescored using Methods 2 and 3 for each of the 10 scoring methods. This was possible since no indication of the termination of the first test was given to the subject and since items were randomly ordered within strata. Once test termination was reached using termination Method 2 or 3, the next item taken by the subject in his entry point stratum acted as the start of a parallel forms test under the termination rule used.

Of course, Method 2 required less items than Method 1 and Method 3 considerably less than Method 2. The thrust of this investigation, then, was to determine the relative efficiency of the three methods in comparison with one another and with linear testing after equalizing test length using the Spearman-Brown prophecy formula.

Stradaptive Test Output

Figure 8 provides an example of a stradaptive test report from this experiment. A "+" next to an item indicates a correct response; a "–" an incorrect response, and "?" shows that the subject omitted the item.

The examinee in Figure 8 estimated her ability as "5." Hence, her first item was the first item in the fifth stratum. She correctly answered this question, but missed her second item, the first item stored in the 6th stratum. She skipped the next item, and after responding somewhat inconsistently for the first nine items, "settled down" with a very consistent pattern for items 10 through 19 when she reached stopping rule criterion and her first test terminated.

At this point in her stradaptive testing, the testing algorithm selected the 6th item in stratum 5 (her ability estimate) to commence her second test. (The subject was totally unaware of this occurrence, as no noticeable time delay occurred between her 19th and 20th items).

At the conclusion of her 31st item, this subject reached termination criterion for her second test, was thanked for her help in this research project, and given her score of 15 correct answers out of 31 questions with a percentage correct of 48.4%.

The scores for this subject are shown for both tests. The interested reader may gain a more thorough understanding of the scoring methods used in this model by tracing this subject's ability estimate scores through Table 3.

IV. PROCEDURES

Item Pool Construction

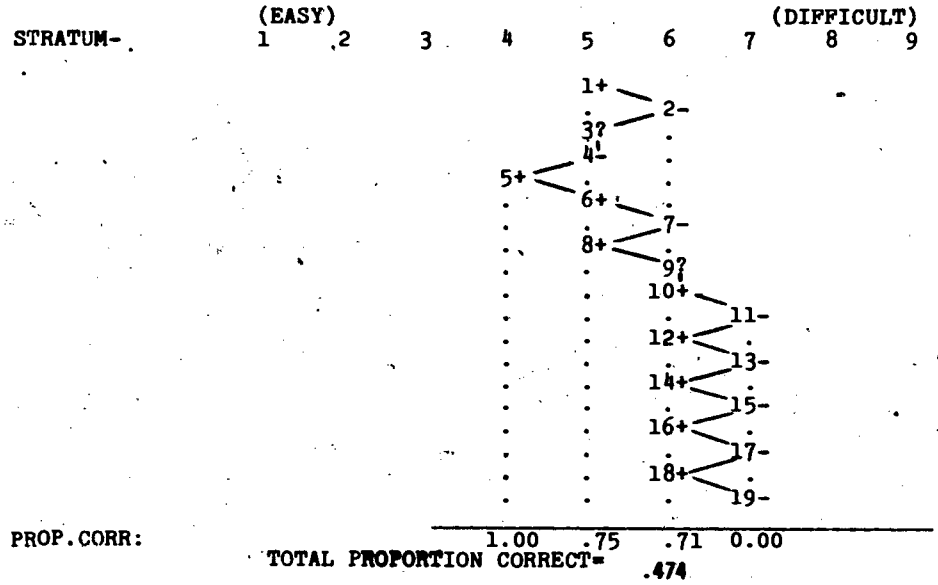
Item pool data received from Educational Testing Service entailed five 50-item verbal analogy tests, Forms 1A, 1B, 1C, 2A and 2B of the SCAT Series II examinations. These tests had been nationally normed on a sample of 3133 twelfth grade students in October, 1966. The five tests were not of equal difficulty, as shown by Table 4, with test 1C considerably more difficult than the other 4 tests.

P-values and biserial correlations were provided by ETS on 249 of the 250 items on the five forms, excluding item number 150, statistics for which were not available. Upon inspection of these indices, item number 169 was removed from the pool due to a biserial correlation of only .10, considered too low for an adaptive test.

Prior to rescaling the item statistics to normal ogive parameters, item difficulties were adjusted by adding an arbitrary value of + .04 to all norm group P-values. This was done to compensate for maturation of subjects between the age at norming and the age at the experimental testing. The SCAT Series I Technical Manual

REPORT ON STRADAPTIVE TEST 1

IDNUMBER- 263354070 DATE TESTED- 74/07/29



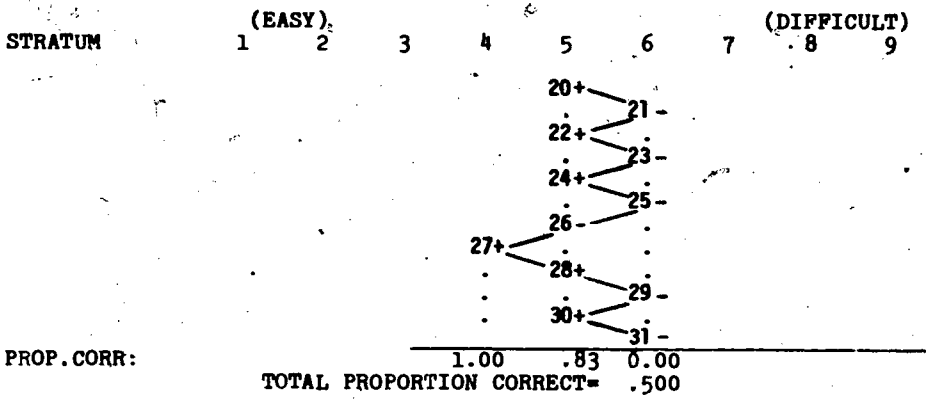
TOTAL PROPORTION CORRECT= .474

SCORES ON STRADAPTIVE TEST 1

1. DIFFICULTY OF MOST DIFFICULT ITEM CORRECT=.24
2. DIFFICULTY OF THE N+1 TH ITEM= .11
3. DIFFICULTY OF HIGHEST NON-CHANCE ITEM CORRECT=.24
4. DIFFICULTY OF HIGHEST STRATUM WITH A CORRECT ANSWER= .04
5. DIFFICULTY OF THE N+1 TH STRATUM=.04
6. DIFFICULTY OF HIGHEST NON-CHANCE STRATUM=.04
7. INTERPOLATED STRATUM DIFFICULTY=.06
8. MEAN DIFFICULTY OF ALL CORRECT ITEMS= -.09
9. MEAN DIFFICULTY OF CORRECT ITEMS BETWEEN CEILING AND BASAL STRATA= -.02
10. MEAN DIFFICULTY OF ITEMS CORRECT AT HIGHEST NON-CHANCE STRATUM= .09

REPORT ON STRADAPTIVE TEST 2

IDNUMBER- 263354070 DATE TESTED- 74/07/29



TOTAL PROPORTION CORRECT= .500

SCORES ON STRADAPTIVE TEST 2

1. DIFFICULTY OF MOST DIFFICULT ITEM CORRECT= -.11
2. DIFFICULTY OF THE N+1 TH ITEM=-.14
3. DIFFICULTY OF HIGHEST NON-CHANCE ITEM CORRECT=-.11
4. DIFFICULTY OF HIGHEST STRATUM WITH A CORRECT ANSWER= -.25
5. DIFFICULTY OF THE N+1 TH STRATUM=-.25
6. DIFFICULTY OF HIGHEST NON-CHANCE STRATUM=-.25
7. INTERPOLATED STRATUM DIFFICULTY=-.18
8. MEAN DIFFICULTY OF ALL CORRECT ITEMS=-.26
9. MEAN DIFFICULTY OF CORRECT ITEMS BETWEEN CEILING AND BASAL STRATA=-.21
10. MEAN DIFFICULTY OF ITEMS CORRECT AT HIGHEST NON-CHANCE STRATUM= -.21

Figure 8. **Example of stradaptive testing report.**

Table 4. **Comparison of SCAT Series II Verbal Forms 1A, 1B, 1C, 2A, & 2B (N = 3,133)**

Form	Item #	Mean	Std Dev	Std Err	KR-20
1A	1– 50	28.7	8.7	3.0	.88
1B	51–100	29.9	8.8	3.0	.88
1C	101–150	24.8	7.5	2.8	.86
2A	151–200	30.5	8.2	3.0	.86
2B	201–250	31.4	8.5	2.9	.88

reported a constant 4% increase in verbal test scores across quartiles between the 12th and 13th grade years. In addition, a restriction of range caused by the selectivity of Florida State University admissions requirements was anticipated, thus making the items for the experimental subjects easier than their normed item parameter estimates.

The difficulty and discrimination indices on the remaining 248 items in the pool were transformed into normal ogive item parameters using the following formulas:

$$P_g = \frac{1}{\sqrt{2\pi}} \int_y^{\infty} e^{\frac{-z^2}{2}} dz$$

$$b_g = \frac{-y}{r}$$

$$a_g = \frac{r'}{(1-r'^2)^{1/2}}$$

where

P_g = the proportion correct for items g

Z = a normal deviate

Y = the inverse of the cumulative normal distribution function at p_g (a normal deviate)

$r = r_{g\theta}$ = biserial correlation of item score and ability

(From McBride & Weiss, 1974)

Appendix B shows the ETS item statistics and transformed normal ogive item parameters. This transformation assumes a normal distribution of ability within the norming group and a metric chosen with mean ability equal to 0.0 and a standard deviation equal to 1.0.

After calculation of the b_g and a_g values four additional items were removed from the stradaptive item pool. Items 101 and 201 had b_g values < -4.00 and items 48 and 250 had b_g values > 4.00. These extreme values were likely outside the ability range of the subject samples and thus would reduce measurement efficiency.

Statistical analysis of the resulting item pool is shown in Table 5. An inspection of Table 5 points out a major problem in the present study. As suggested in Chapter III of this paper, a restriction of range was anticipated due to the selectivity of Florida State University admissions. In addition the mean difficulty index of –.368 reflects an item pool somewhat too easy (most likely a result of the .04 increase in P_g values).

Table 5. **Descriptive Statistics of Difficulty (b_g) and Discrimination (a_g) Normal Ogive Parameter**

Normal Ogive Parameter	Mean	Std Dev	Std Err	Kurtosis	Skewness
Discrimination (a_g)	.576	.175	.011	−.19	.37
Difficulty (b_g)	−.368	1.132	.072	.33	.37

The distribution of a_g values was satisfactory, with only a slight skew and a mean s_g of .57, but the combined effect of a relatively easy item pool coupled with an expected high ability subject pool suggested the possibility of lowered validity, and internal consistency reliability coefficients for the conventional (linear) test group.

Subject Pool

Each summer, approximately six weeks prior to the start of the academic school year, Florida State University conducts a three-day University orientation for incoming freshmen. In late July, 1974, thirteen hundred students attended the orientation program, 27% of the scheduled first year enrolees.

Each orientation participant received welcoming packages including a letter from the author of this paper. Appendix C presents a copy of the letter, which requested voluntary participation in a computer-based research project. The voluntary nature of the request was required by University orientation officials. One other source of subject recruitment was utilized. The CRT's used in this experiment were located in the FSU library's listening and viewing center. The library held three library orientation tours each day of the orientation program to acquaint the new students with the library facilities. When these groups were brought to the area of the listening and viewing center, the author of this paper made "a pitch" for volunteers for the project.

Of the 103 subjects who participated in the research, 87 had previously taken the Florida 12th Grade Verbal Ability test (12V). Like the SCAT-V, items used in the item pool, 12V items were verbal analogies, prepared by ETS for the State of Florida. Item format was identical to SCAT Series II Verbal item format. Reliability (KR-20) of the 12V was reported as .87 for 50 items with a 20 minute time limit. The 12V, thus, provided an ideal validity criterion for comparison with linear and stradaptive scores from this experiment. In addition, 12 of the subjects without 12V scores had taken either the American College Testing Program (ACT) or College Entrance Examination Board (CEEB) verbal tests which had equivalency tables to the 12V. No criterion scores were available for two of the stradaptive subjects and one of the linear examinees. Validity indices were thus computed with N = 53 for the stradaptive group and 46 for the linear.

Table 6 shows the comparison between 12V norming group statistics and the subjects sampled in this experiment. As can be readily seen in Table 6, the suspected restriction of range was certainly evident.

Table 6. **Comparison of Florida 12th Grade Verbal Test Scores (1973 Statewide Administration vs. Subject Sample)**

Test Group	N	Mean 12V Score	Std Dev
Statewide Norming Group	81000	26.15	8.26
Experiment Participants	99	33.83	5.94

Pr (μ state = μ exp) = < .001
Pr (σ^2 state = σ^2 exp) = <.001

Both means and variances of 12V scores are significantly different from those of the population, with the restricted variance of participants in this study predominantly caused by admissions policies, but also possibly by a "ceiling effect." Regardless of cause, this restriction would lower validity indices within the relatively homogeneous group of subjects in this experiment.

Fortunately, the primary comparisons of interest in this study were between the stradaptive and linear test group participants. Table 7 shows the comparison between these two groups within the experiment.

Table 7. **Comparison of Distributions of Linear and Stradaptive Group Florida 12th Grade Verbal Scores**

Group	# Subjects	Mean	Std Dev	Std Err	Kurtosis	Skewness
Linear	46	33.26	5.30	.855	.44	.70
Stradaptive	53	34.06	6.12	.841	.36	−.03

$P_r(\mu_{lin} = \mu_{str}) = .05$
$P_r(\sigma^2_{lin} = \sigma^2_{str}) = .05$

As can be seen in Table 7, the random assignment of subjects to linear or stradaptive testing groups did a good job of equating the groups on the ability continuum as measured by the Florida 12th Grade Verbal test.

Research Design

Prior to data collection, 300 random assignments were made to either linear or stradaptive groups and the linear group was further randomly broken into five subgroups corresponding to the five linear subtests.

As subject-volunteer entered the testing area, the proctor assigned him the next test listed on the randomized testing order schedule. Schematically, the research design is depicted in Figure 9.

A comparison of outcomes 0_1 through 0_5 would indicate the effectiveness of the randomization process in equating subtest assignment. Assuming no significant differences between these outcomes, comparisons between 0_6 through 0_{10} could then be made. Since SCAT-V published results had shown significantly different difficulty levels between the five forms, it was planned that linear subtest scores would be normalized within their separate distributions and then pooled into a linear total score distribution for comparison with the stradaptive results.

The independent variables for the comparisons in this study were linear or stradaptive group, termination rule, 12V score and scoring method. Dependent variables included test scores, item latency, number of items, standard errors (and/or reliability), and validity.

Data Collection

A file was created as each subject went through the instructional and testing process. A description of data collected is listed in Appendix D. Item data stored included response code (correct, incorrect or skipped), the subject's actual response (A, B, C, D or ?), the number of the item in the total pool (1–250), the number of presentations of the question, and item response latency in seconds. This data was collected for each of a maximum of 60 items, with the word "break" inserted in the item data file between the first and second tests of a stradaptive subject.

These subject data files were stored separately under individual file names for later analysis and computer-generated reports like Figure 8.

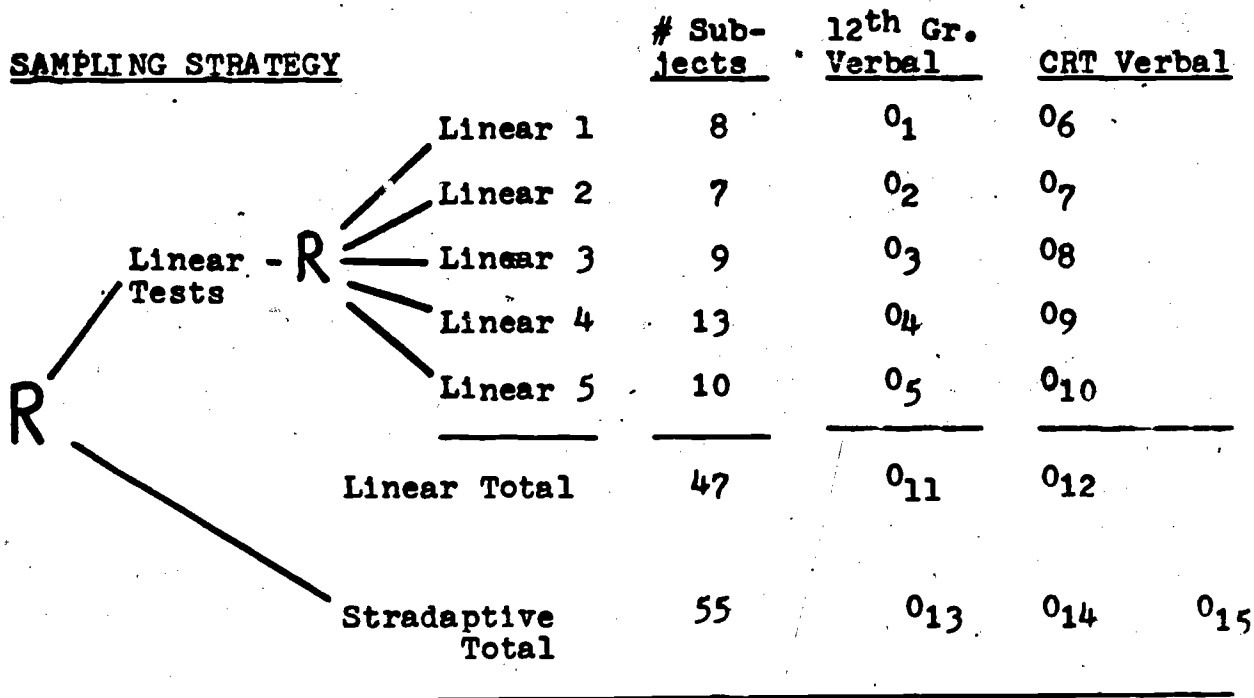


Figure 9. **Research design for linear versus stradaptive group assignment and comparison.**

Data Analysis

The following analyses were planned:

1. Total linear vs. total stradaptive using 3 termination rules and 10 scoring methods.
   (a) Standard errors of measurement
   (b) Reliability (parallel forms and KR-20)
   (c) Validity (correlation between 12V and test scores) number of items per terminated test
   (d) Item latency
2. Correlation between the linear subject's ability estimate and his 12V score and linear test score.
3. Correlation between the linear subject's 12V scores and item latency.
4. Correlation between scores of any subjects who took both linear and stradaptive tests. (This situation was not part of the original design of this experiment, but a few subjects requested to "do it again" and were administered the "other" test). This correlation coefficient would be spuriously high due to common items between the linear test and approximately $1/5^{th}$ of the items on the stradaptive test.

Attitudinal Data

Consideration had been given to preparing a questionnaire to survey subject reaction to the computer-based mode of test presentation used in this study. It was decided to forego a formal attitudinal study for the following reasons:

1. Considerable evidence already exists pertaining to subject reaction to computer-assisted testing and instruction (Hansen, 1969). The computer mode of presentation evidently does not decrease subject test performance.

2. The main thrust of the current research was a validation of the stradaptive model, not of computer-testing.

3. Subjects took only computer-based testing and therefore probably had no realistic basis of comparison.

Despite these considerations, the closing screen shown each subject before he left the CRT did request any comments he might have about computer testing "to aid the researchers in future studies." These comments were jotted into a ledger for synopsizing in the conclusion section of this paper.

V. RESULTS AND DISCUSSION

Table 8 shows a comparison of the distribution of the five linear subtests and their respective 12V score distribution.

Table 8. **Comparison of Distributions of 5 Linear Subtests**

Subtest Group	12th Grade Score			. Subtest Score			
	N	Mean	Std Dev	P_c	Std Dev	Kurtosis	Skewness
1	8	36.1	7.43	.76	.11	−.69	.32
2	7	32.6	3.82	.68	.15	−.96	.52
3	9	30.5	3.62	.53	.08	−1.39	.24
4	13	33.4	6.65	.81	.08	−.61	−.56
5	10	32.4	4.67	.76	.10	−.47	−.56

Surprisingly, the mean 12V score of the group taking linear test 1 was significantly higher than the other four groups ($p =< .05$). In the comparison of the proportion of the items answered correctly (omits counted wrong) on the five subtests, linear 4 was significantly easier than linear 2 and as expected, linear 3 was significantly more difficult than the other four subtests. In addition, linear 3 produced a decidedly platykurtic distribution, while linear 4 and linear 5 evidenced a concentration of responses at the higher end of the distribution.

Despite these differences in distribution shape, the five subtests were normalized and then pooled for group comparison with stradaptive test results. The resulting distribution of total linear scores is shown in Table 9. The distribution was essentially normal, though platykurtic.

Table 9. **Distribution of Pooled Linear Test Scores**

Mean	Std Dev	Std Err	Kurtosis	Skewness
−.02	.946	.138	−.67	0.06

Linear Test Reliability

Stanley (1971) described the procedures for estimating the internal consistency reliability (KR-20) for a test in which different subjects took different items and different numbers of items from a unidimensional pool.

Making the standard assumptions underlying the one-factor random effects analysis of variance (ANOVA), an estimated reliability coefficient of the total scores, Xp, of persons receiving l_p items may be obtained through the use of the following formula:

$$\rho^2_{TX} = \frac{l_p \rho_{intraclass}}{1 + (l_p - 1)\,\rho_{intraclass}} = 1 - \frac{MS_e}{MS_a} \qquad (11)$$

Table 10 displays the ANOVA source table for the linear group in this experiment. The internal consistency reliability estimate for the linear test was .776 for a test of an average of 48.4 items in length. Stepped-up to 50 items via the Spearman-Brown Prophecy formula, this estimate becomes .782. The comparable reliability of the original SCAT-V tests was .87. Using Feldt's (1965) test, $Pr(\rho_{scat} = \rho_{lin}) = <.05$.

Table 10. **Analysis of Variance for Linear Test Person by Item Matrix**

Source	df	Sum of Squares	Mean Squares
Persons	46	37.57	.817
Error	2229	408.55	.183
Total	2275	446.12	

$$r_{tx\,(lin)} = 1 - \frac{.183}{.817} = .776$$

It can be assumed that the difference these reliabilities was caused by one or more of three factors:

1. Testing mode (CRT vs. paper and pencil)
2. Elimination of 6 items from the original item pool.
3. Restriction of range in subject pool for this experiment.

The latter factor most likely caused the majority of the decrease in the reliability of the test scores. The homogeneity of the subjects would yield a relatively small amount of between-person variance, which, when coupled with a constant error variance, would lower the reliability estimate. It might also be mentioned that Stanley noted that intraclass item correlation is a lower bound to the reliability of the average item.

Test theory suggests that measurement efficiency is maximized at p = .50 for a given test group. It was hypothesized that the stradaptive test strategy would better approach this standard than the conventional linear test. If supported, this result would indicate an improved selection of items for the stradaptive examinee. Table 11 shows the result of this comparison. It clearly indicates significantly different distributions of test difficulty. The stradaptive test was far more difficult than the linear test, with a smaller variance.

Table 11. **Comparison of Difficulty Distributions (P_c) for Linear and Stradaptive Groups**

Group	#Subjects	(P_c)	Std Dev	Std Err	Kurtosis	Skewness
Linear	47	.752*	.123**	.018	−.87	−.39
Stradaptive	55	.584	.084	.011	5.14	1.97

*Pr ($\mu_{Str} = \mu_{Lin}$) = <.0001
**Pr ($\sigma^2_{Str} = \sigma^2_{Lin}$) = <.05

$$*Z = \frac{(X_a - X_b) - (\mu_a - \mu_b)}{\sqrt{(\sigma_a^2/n_a) + (\sigma_b^2/n_b)}}$$

This test makes no assumption about the equality of population variances. (from: Winer, 1971)

$$**C = \frac{S^2 \text{ largest}}{\S S_j^2}$$

Cochran's Test for Homogeneity of Variance (from: Winer, 1971)

Linear Test Validity

The reported correlations of the SCAT-V Series II scores with several criteria are summarized in Table 12. The correlation of obtained linear scores with the Florida 12th Grade scores was .477, which was lower than the published SCAT-V:SAT-V correlation (p = <.01). As with the linear reliability, this difference probably resulted from the homogeneous distribution of subjects in this experiment.

Table 12. **Reported Correlations of SCAT-V Scores with External Criteria**

Criterion	N	r_{12}
High School English Grades	244	.46
Normalized Rank in Graduating Class	244	.49
Rank in Graduating Class	518	.52
SAT-V	244	.83

Stradaptive Pool Item Stratification

Table 13 summarizes the proportion of items in each stratum that were actually used in the stradaptive testing.

Table 13. **Proportion of Items in Each Stratum Actually used in CRT Stradaptive Testing (N = 55)**

	Stratum								
Proportion	1	2	3	4	5	6	7	8	9
Number of Items in Stratum	20	26	33	39	31	28	26	22	19
Available Items Used Within Stratum	.10	.12	.18	.38	.68	.61	1.00	1.00	1.00

The results depicted in Table 13 tend to contradict Weiss' suggestion that a larger proportion of items should be assigned to lower and middle strata (Weiss, 1973). The present author recommends that the decision be based upon prior knowledge of the distribution of ability of the subject pool to be tested. Such prior knowledge includes school admissions requirements and any other information the decision-maker may have available about the target population ability level.

Stradaptive Total-Test Reliability

Using Stanley's (1971) procedure, it was possible to estimate the internal-consistency reliability of the person-by-item stradaptive test matrix using scoring method 8. Appendix A, columns 7–9, shows the pattern of item presentation across subjects. Of the 244 items in the stradaptive pool, only 133 items were actually presented to the subject pool in this experiment.

Scoring method 8 provided the only set of stradaptive test scores wherein a person's total test score was a linear function of his item scores. Hence, scoring method 8 was used to estimate internal-consistency reliability using Stanley's ANOVA procedure. Table 14 summarizes these results.

In addition to the internal-consistency reliability estimate shown in Table 14, parallel-forms correlation on the total stradaptive pool using the three termination rules with ten scoring methods were calculated. Table 15 displays these results.

Table 14. **Analysis of Variance of Scoring Method 8 of Stradaptive Test Person-by-Item Matrix**

Termination Rule	Source	df	Sum of Squares	Mean Squares
1	Persons	54	191.941	3.555
	Error	1675	588.253	.351
	Total	1729		(r_{20} = .901)
2	Persons	54	178.870	3.312
	Error	1401	470.442	.336
	Total	1455		(r_{20} = .899)
3	Persons	54	155.841	2.886
	Error	1001	366.447	.366
	Total	1055		(r_{20} = .873)

Table 15 shows the statistical analysis of the differences between parallel-forms reliability estimates on the stradaptive test scores. Significance of the differences in reliability coefficients (r_{xx}) was determined using Ferguson's (1971) formula.

Table 16 shows the parallel-forms and KR-20 reliability estimates for the three termination rules used in this study. Direct comparisons can be made between the stradaptive KR-20 values and the .776 linear KR-20 estimate. According to Feldt's (1965) approximation of the distribution of KR-20, all of the estimates of the stradaptive test reliability are significantly ($p = <.05$) better than the linear KR-20 estimate prior to being stepped-up by the Spearman-Brown formula $Pr(.675 < P_{20} < .858) = .95$. Thus, the 19, 26, and 31 item stradaptive tests all proved more reliable than the 48 item linear test. This is the key finding in this study.

A comparison of the linear internal-consistency reliability coefficient (r_{tx}) and the stradaptive parallel-forms reliability estimate (r_{xx}) can be considered only tentatively since they are a different kind of estimate of the true reliability. The sampling distribution of r_{xx} is known and that of r_{tx} has been approximated by Feldt (1965). Cleary and Linn (1969) compared standard errors of both indices with

TABLE 15

Comparison of Parallel-Forms Reliabilities for 10 Stradaptive Test Scoring Methods under Three Termination Rules Stepped-Up to 50 Items

TERMINATION RULE 1

(N = 12)

SCORING METHOD	8	6	7	9	4	10	3	1	5	2
($\bar{I}_p$ = 31.45) r_{xx}	.929	.910	.902	.879	.703	.620	.616	.436	---[1]	---[1]
Statistically Significant differences	*						*			

TERMINATION RULE 2

(N = 28)

SCORING METHOD	8	9	7	6	3	10	1	4	5	2
($\bar{I}_p$ = 26.47) r_{xx}	.806	.782	.750	.698	.682	.614	.432	.379	---[1]	---[1]
Statistically Significant differences	*						*			

TERMINATION RULE 3

(N = 38)

SCORING METHOD	8	6	7	3	10	9	5	1	2	4
($\bar{I}_p$ = 19.2) r_{xx}	.903	.821	.820	.791	.784	.689	.590	.587	.582	.513
Statistically Significant differences	*					*				

[1] negative parallel-forms correlation - differences not calculated

$\bar{I}_p$ = mean number of items for this termination rule

* $p = < .05$ between ⊢——⊣'s

Table 16. **Comparison of Scoring Method 8 Parallel Form Reliability with KR-20 Reliability Over Three Termination Rules Stepped Up to 50 Items**

		Termination Rule 1	2	3
		(N = 12)	(N = 28)	(N = 38)
Parallel Forms	r_{xx}(raw)	.892	.688	.732
	r_{xx}(50)	.929	.806	.903
		(N = 55)	(N = 55)	(N = 55)
KR-20	p_{20}(raw)	.901	.899	.873
	p_{20}(50)	.935	.943	.947
		K_i = 31.45	K_i = 26.47	K_i = 19.2

K_i = average number of items under termination rule i.

generated data of known ρ. They found the standard error of KR-20 to be somewhat smaller than that of the parallel-test correlation (approximately .05 vs. .04 in the range of reliabilities, number of subjects, and number of items involved in this experiment). Should these results generalize to this study, scoring methods 6, 7, 8, and 9 under termination rule 1, and scoring method 8 under termination rule 3 produced higher reliability than the linear test.

The interpretation of the results shown in Table 15 was clear. In the comparison of scoring methods, methods 6, 7, 8, and 10 were significantly (α = .05) more reliable than methods 1, 2, and 5 within all three termination rules. Scoring method 8 produced the highest reliability estimate under all three termination rules. In the comparison between the three termination rules, methods 1 and 3 are significantly better than method 2 ($p = <.05$) using the Wilcoxon Matched-Pairs-Signed-Ranks Test (Siegel, 1956).

Stradaptive Test Validity

The validity coefficients of the 10 stradaptive scoring methods under the three termination rules is shown in Table 17. Validity was estimated by the correlation between the test scores and 12V scores.

Table 17. **Comparison of Validity Coefficients of 10 Stradaptive Test Scoring Methods Under Three Termination Rules**

				Termination Rule 1 (N = 64)						
Scoring Method	8	9	1	5	7	3	10	6	2	4
r_{c1}	.526	.513	.477	.443	.437	.425	.395	.385	.380	.370
				Termination Rule 2 (N = 80)						
Scoring Method	8	9	7	3	5	1	10	6	2	4
r_{c2}	.536	.501	.471	.420	.403	.397	.393	.365	.350	.275
				Termination Rule 3 (N = 91)						
Scoring Method	7	5	8	3	9	6	2	10	1	4
r_{c3}	.509	.500	.499	.492	.476	.467	.455	.442	.410	.240

r_{ci} – correlation between criterion measure (12V) and scoring method i.

Among the ten scoring method validity coefficients, the following comparisons showed significant differences ($p = <.05$):

Termination method 2:

Scoring method 8 > scoring method 4.

Termination method 3:

Scoring method 7, 5, and 8 > scoring method 4.

None of the validity coefficients in Table 16 were significantly different from the linear validity coefficient of *.477*. Since scoring methods 6, 7, and 8 and 10 were consistently more reliable than the other methods, the validity coefficients for these four methods were raised by the so-called "correction for attenuation" for comparison purposes. Table 18 shows the results of this adjustment.

Table 18. **Effect of the Four Most Reliable Stradaptive Scoring Methods Correlation with 12V, Corrected for Attenuation**

Termination Rule		Scoring Rule 6	7	8	10
1	r_{xx}	.910	.902	.929	.620
	r_{xc}	.385	.437	.526	.395
	(r_{xc})	(.433)	(.493)	(.585)	(.538)
2	r_{xx}	.698	.750	.806	.614
	r_{xc}	.365	.421	.536	.393
	(r_{xc})	(.528)	(.544)	(.693)	(.623)
3	r_{xx}	.821	.840	.903	.784
	r_{xc}	.467	.509	.499	.442
	(r_{xc})	(.627)	(.684)	(.626)	(.621)

r_{xx} = parallel forms reliability estimate
r_{xc} = correlation of scoring Method Test Score with 12V
(r_{xc}) = r_{xc} corrected for unreliability of 12V and stradaptive scoring method

When both validity and reliability were considered, stradaptive scoring methods 7 and 8 were judged superior to the other methods considered in this study.

Method 8, the mean difficulty of all items answered correctly, has several characteristics to recommend it. It would seem to use the maximum amount of information available from the subject's responses. Since the subject's total score under method 8 is a linear transformation (a mean) of the item scores, Stanley's (1971) ANOVA internal-consistency reliability estimating procedure is applicable. For both experimental and applied situations, a single testing design is more feasible than a test-retest or parallel-forms design.

Method 8 does suffer from two conceptual flaws. Whenever a subject's ability estimate (entry point) was grossly low, scoring method 8 would be biased toward a lower estimate of the subject's true score. In addition, the method is inflated by "lucky guessing." If an ability estimate were prestored on subjects or if it could be assumed that they could estimate their own ability fairly well, method 8 would be the best method of implementing stradaptive testing.

In the present study, the correlation between the subjects' ability estimates and their total linear score was *.466*, essentially as good a predictor of their linear scores as the Florida 12th Grade Verbal test scores *(.477)*. Under such a situation, scoring method 8 appears to be conceptually sound as an estimator of a subject's true score.

In the case where no ability estimate was available for examinees, and it could not be assumed that they could fairly accurately estimate their own ability (young children, for example) method 7 would be the recommended scoring method on a stradaptive test.

Linear vs. Stradaptive Comparisons

Given the stradaptive-test scoring recommendations in the previous section, how do linear and stradaptive testing procedures compare overall? Table 19 shows the results of the three termination rules for scoring method 8 of the stradaptive test along with linear test statistics.

Table 19. **Comparison of Linear Test with Scoring Method 8 Under Three Termination Rules of the Stradaptive Test**

	Stradaptive Test Termination Method		
Linear Test	**1**	**2**	**3**
	Total Test Variance		
.817	.403	.433	.433
	Standard Error of Measurement		
.428	.162	.157	.157
	KR-20 Reliability		
.776	.935	.943	.947
	Parallel Form Reliability		
*	.929	.806	.903
	Validity		
.477	.526	.536	.509
	Validity (Corrected for Attenuation)		
.546	.585	.693	.626

*No linear parallel-forms reliability calculated.

Table 19 provides strong evidence that the measurement efficiency of the average item on the stradaptive test is as good or better than the conventional test. Nevertheless, unless a reduction in the number of items required occurs, as well as a reduction in testing time, the theoretical gain in efficiency may not have real-world value.

Table 20 shows the difference in number of items presented for the linear and the three termination methods of the stradaptive test. The consistency in average number of items presented per subject was surprisingly constant over the two parallel tests of termination methods 1 and 3. Method 2 did show a

Table 20. **Comparison of Average Number of Items for Linear Test and Three Termination Methods of Alternate Form Stradaptive Tests**

Test	# Subjects	Avg # Items	Std Dev # Items		Avg # Items	Std Dev # Items
Linear	47	48.43	.99			
Stradaptive		**Test 1**			**Test 2**	
Method 1	55	31.46	18.03	38	30.92	12.54
Method 2	55	26.94	16.76	41	21.98	13.10
Method 3	55	19.20	14.06	47	18.19	11.34

significant ($p =< .05$) drop in the average number of items on the second test, possibly due to the 60-item limit.

It was hypothesized that mean latency would be higher for stradaptive subjects since they would have to "think" about each item as it was near the limit of their ability. Table 21 reflects the results of this comparison.

Table 21. **Comparison of Distributions of Item Latency Between Linear and Stradaptive Groups**

Group	Items	Mean Sec/Item	Std Dev
Linear	2276	35.999*	12.062*
Stradaptive	1730	40.047	13.219

$\Pr(\mu_{str} = \mu_{lin}) =< .001$
$\Pr(\sigma^2_{str} = \sigma^2_{lin}) =< .001$

The hypothesis of no differences between item latencies was rejected. For the subjects in this experiment, the average stradaptive item required approximately 11% longer than the average linear item.

Omitting Tendency

The analysis of the relationship between the tendency to omit and ability was investigated. If the hypothesis of no differences in the tendency across ability levels was rejected the handling of the omits could create a bias in total test scores. For the subjects in this experiment, the correlation between omitting and 12V score was $-.07$, $P_r(r_{omit/12V} \neq 0) > .05$, thus the hypothesis of no difference was not rejected.

Correlation Between Scores of Subjects Who Took Both Stradaptive and Linear Tests

Six subjects asked to return the next day and take "the other" test. This second testing was permitted, with the resulting test score data withheld from analysis except for this section of the paper. The correlation between the scores of subjects on both testing strategies provides an indication of the unidimensionality of the underlying psychological trait common between the two tests. It must be kept in mind, however, that the stradaptive item pool was made up of items from the five linear subtests. Thus a dependency between test methods existed. It would be expected that approximately 1/5th of the items taken on the stradaptive test also appeared on a subject's linear test. The standardized linear scores and stradaptive score 8 counterparts are shown in Table 22. Correlation between the measures was .93.

Table 22. **Linear and Stradaptive Scores of Subjects Who Took Both Tests**

Subject	Linear	Stradaptive
1	.82	.67
2	−.06	.30
3	−.14	−.23
4	.68	.81
5	.83	.76
6	−.25	−.16

$r_{12} = .931$

Attitudinal Information

The overwhelming proportion of comments received after the testing was favorable to computer-based testing. Only one subject reported prior experience with CRT operation, yet no major problems arose in any students operating the equipment.

Stradaptive subjects tended to comment that the test was "very hard" and some expressed anxiety at only getting about half of the items right. This problem suggests that perhaps adaptive testing subjects should be led to anticipate "only" getting 50% of the items correct in order to keep student motivation up.

The general reaction of the linear subjects was that the test was "a snap," which was consistent with the over 75% correct response rate shown by the linear group.

Testing Costs

No full cost analysis was planned for this study. However, computer costs were available for the three-day data collection. A total of *$89.00* was spent over the entire period. This total included core memory (CM), central processor (CP), permanent file storage (MS), data transmittal between the CRT's and the computer, line printing (LP), and punch card output for 109 subjects. The author had data files punched-out as they were created to assure that data would not be lost in case of a hardware malfunction.

The cost of testing each individual came to less than 2 cents per subject for CM, CP, MS, and LP time on the CDC 6500 computer. Excluding software preparation costs and hardware rental, etc., this is the expected computer cost per subject in a large-scale testing program, once set up and operating. The salary of proctors has not been included in this analysis, although this cost would certainly be small when pro-rated over a large number of subjects.

In the present study, 6 CRT's were kept on and tied to the computer continuously for 14 hours a day for 3 days in order to be ready for subject-volunteers whenever they arrived. In any implementation of computer-testing outside the experimental situation, exam time would be scheduled, thus minimizing telephone line transmittal costs.

This cost approximation could be compared with testing costs from the reader's experience. Without trying to define conventional test cost per se, there is still little doubt that computer-based testing costs less than conventional testing with the paper and pencil mode for any large-scale testing program.

V. CONCLUSIONS AND IMPLICATIONS FOR FUTURE RESEARCH

The results of this study favor the further investigation of the stradaptive testing model. The model produced validity coefficients comparable to conventional testing with a reduction of the number of items from 48 to 31, 25 and 19 for the three stradaptive termination rules investigated in the study. The internal consistency reliability for the best stradaptive scoring method was significantly higher than the conventional KR-20 estimate, and the stradaptive parallel-forms reliability estimates were consistently higher than the conventional KR-20 for the best of the scoring methods.

The author was not aware of any prior research showing a comparison of item latency data between adaptive and conventional testing modes. Results in this study clearly indicate that subjects take significantly longer to answer items adapted to their ability level, about 11% longer in the present study. This is an important result, as it indicates that future research into adaptive testing of any kind should take this variable into consideration when evaluating an adaptive test strategy. The net gain of the adaptive model is really a function of the testing time needed to adequately measure a subject's ability, not the number of items presented to the subject. All prior research reviewed tacitly assumed that item latency was consistent across testing strategies. This study indicated this assumption is false.

The statistical power of the tests for significant difference between the experimental and control groups in this study was too low. Nearly every researcher is forced to "settle" on a smaller "n" than desired due to the external constraints imposed on his research. This was certainly true in the present study. It is the author's

intent to make this study the first step in an on-going investigation of the stradaptive model, much as Weiss is doing at the University of Minnesota. Where significant differences did not emerge, as in the validity coefficients, the trend was consistently favorable to the best stradaptive models in comparison to the linear models. Should this trend be upheld as the number of subjects in the research grows, stronger statements about the comparative validities of the two methods could be made. This possibility alone suggests that model investigation be continued.

Within the three termination rules investigated, KR-20 reliabilities were essentially equal for a test length of 50 items. Termination method 3, however, would have yielded an equivalent reliability estimate at 25 items to the "raw" KR-20 estimates of the other two methods at 26 and 31 items. This result supports Novick (1969) and Wood (1971) evidence that the efficiency of adaptive testing "levels off." Their result on Bayesian models suggested that from 15 to 20 items was optimal, as opposed to the 25-item "peak" shown in the present study.

The validity comparisons between the three termination strategies did not yield significant differences. The trend, however, consistently showed method 2, wherein omits were counted as wrong and 5, the minimum number of items in the ceiling stratum, as producing poorer measurement than the other two termination methods. This result is difficult to explain. Method 1 ignored omitted items and set the minimum number of items in the ceiling stratum to 5. Method 2 considered omits wrong, but used the same test termination rule for the ceiling strata. Theoretically, the consistent difference between these two methods should reflect that the first treatment of omits was better. Method 3, which used an identical treatment of omits to method 2, but set the stopping minimum at 4 items in the ceiling stratum, was also better than method 2. This second result suggests that presenting *less* items yields higher reliability when omits are counted as incorrect answers.

The analysis of the termination rule is further complicated by the existence of Weiss' other branching model. In the present author's judgment, the strategy of branching to a lower stratum after an omitted item is conceptually superior to the repetition of another item within the same stratum. Weiss' preliminary results (personal communication) support this hypothesis as he has consistently found the test-retest reliability of the first model to be about .10 higher than the model used in this present study. Given the model evaluated in this experiment, the author would recommend that termination method 3 be used in future stradaptive testing since its measurement effectiveness is comparable to the other 2 methods, but with less items.

In the comparisons of scoring methods, the mean difficulty of all items answered correctly is recommended for any subject pool whom it could be assumed would adequately estimate their own ability. Scoring methods 6, 7, and 10 yielded parallel-forms correlations that were statistically equivalent to method 8, but methods 6 and 10 consistently produced lower validity. These results are understandable for method 6, the mean difficulty of all items in the highest non-chance stratum. The author would expect this estimate to be fairly accurate, but unfortunately the number of possible scores using methods 4, 5, and 6 is limited to the number of strata in the item pool. Method 7, the interpolated stratum difficulty, corrects for this deficiency in method 6. Method 10, the average difficulty of the correct responses in the highest non-chance stratum, is conceptually appealing to the present author. The ability estimate from scoring method 10 is not affected by a poor entry point ability estimate by the subject or by "lucky" guesses about a subject's ability stratum.

It is recommended therefore, that future stradaptive experimental studies concentrate upon scoring methods 7, 8, and 10. These studies should also consider both stradaptive branching models with a comparison of results from variation in the minimum number of items in the ceiling stratum. A comparison between these variable number of stage strategies and several fixed number of stage strategies is desirable. The author plans such an analysis on the present data in the near future. As suggested in prior research, adaptive testing may reach "peak" efficiency at between 15 and 20 items. A comparison of stradaptive test statistics for example with k = 10, 15, 20, and 25 items with linear testing should investigate this hypothesis. Once the stradaptive data is collected under the variable strategy, the fixed item statistics can be determined by grading the stradaptive test after "K" items and then "starting" the subject's second test at the first item of the entry point level.

One further suggestion for future stradaptive studies has occurred to the author. Following the same logic which led to termination of a subject's testing when 5 items in a row in the highest stratum had been correctly answered, the missing of 5 items in a row in *any* stratum should provide an immediate ceiling stratum definition. The probability of the occurrence would be less than .05 for a properly normed item pool. In the case of the present study, 13 of the 55 stradaptive subjects would have terminated in a stradaptive test an average of *12.1* items earlier than termination method 1, with no effect upon the other 42 subjects. The resulting stradaptive test statistics obtained from the implementation of this suggestion have not been calculated, except that the change would have reduced the average number of items presented under termination method 1 to 28.4 from 31.45. The author plans this test statistic analysis for the near future. However, the suggestion was listed here for the consideration of any other stradaptive investigators.

Aside from the stradaptive model per se, further research into adaptive testing in which both the number of stages and step-size are variable is recommended. The Bayesian strategies and Urry's model are examples of this category of adaptive measurement, and further model development seems appropriate.

Research is necessary with comparisons between stradaptive models rather than the traditional design of comparing adaptive method with the conventional method of testing. Weiss' on-going research project is beginning this type of work, but more is needed. The traditional comparison assumes that conventional test statistics are the criterion that an experimental testing procedure should try to duplicate. Lord, Green, Weiss et al., have argued that improved measurement of the individual at all ability levels may be hidden by the use of classical test statistics such as validity and even reliability. Levine and Lord (1959) suggested an index of discrimination which considered various levels of the test score range and Lord's (1972) information function theory and item characteristic curve theory are an attempt to solve this problem. More theoretical research in this area is needed.

The goal of this study included the attempt to estimate the degree to which the violation of the assumptions of the one-factor ANOVA model affected KR-20 reliability estimates. The assumption that items are independent of one–another clearly is violated in any adaptive testing procedure. The degree of effect this assumption violation causes is unknown, yet most prior research in adaptive testing which has considered reliability at all, has only considered ANOVA KR-20 estimates.

Certainly the results from this study do not allow any definitive statements about this question. Nevertheless, the three KR-20 estimates were consistently higher than the 3 parallel-forms reliabilities. Cleary and Linn's (1969) monte carlo study indicated that r_{20} provided better parameter estimation than parallel-forms reliability estimates, so one must question whether the higher ρ estimates are not the result of the dependency between items. Perhaps the only way this question can be answered is through a monte carlo study of adaptive testing with ρ known and the two methods compared, for estimating ρ.

Green (1970) concluded that the computer has only begun to enter the testing business, and that as experience with computer-controlled testing grows, important changes in the technology of testing will occur. He predicted that "most of these changes lie in the future. . . . in the inevitable computer conquest of testing."[3]

The stradaptive testing model would appear to be one such important change.

---

[3]Green, B.F., Jr. In Holtzman (Ed.), 1970, p. 194.

REFERENCES

Anastasi, A. An empirical study of the applicability of sequential analysis to item selection. *Educational and Psychological Measurement,* 1953, 13, 3–13.

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

Armitage, P. Sequential analysis with more than two alternative hypotheses, and its relation to discrimant function analysis. *Journal of the Royal Statistical Society*, 1950, Vol. 12, 137–144.

Baker, F.B. An intersection of test score interpretation and item analysis. *Journal of Educational Measurements*, 1964, 1, 23–28.

Bayroff, A.G. *Feasibility of a programmed testing machine.* U.S. Army Personnel Research Office Research Study 64-03, November 1964.

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

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

Betz, N.E., & Weiss, D.J. An empirical study of computer-administered two-stage ability testing. *Research Report 73-4,* Psychometric Methods Program, University of Minnesota, 1973.

Boldt, R.G. Study of linearity and homoscedasticity of test scores in the change range. *Educational and Psychological Measurement,* 1968, 28, 47–60.

Bryson, R. *A comparison of four methods of selecting items for computer-assisted testing.* Technical Bulletin STB 72-5, Naval Personnel and Training Research Laboratory, San Diego, December 1971.

Bryson, R. Shortening tests: effects of method used, length and internal consistency on correlation with total score. *Proceedings, 80th annual convention of the American Psychological Association,* 1972, 7–8.

Buros, O.K. *The 7th mental measurements yearbook.* Highland Park, N.J.: Gryphon Press, 1971.

Cleary, T.A., & Linn, R.L. A note on the relative sizes of the standard errors of two reliability estimates. *Journal of Educational Measurement,* 1969, 6, #1, 25–27.

Cleary, T.S., Linn, R.L., & Rock, D.A. An exploratory study of programmed tests. *Educational and Psychological Measurement,* 1969, 28, 345–360. (a)

Cleary, T.A., Linn, R.L., & Rock, D.A. Reproduction of total test score through the use of sequential programmed tests. *Journal of Educational Measurement,* 1968, 5, 183–187. (b)

Cowden, D.J. An application of sequential sampling to testing students. *Journal of the American Statistical Association,* 1946, 41, 547–556.

Cronbach, L.J. New light on test strategy from decision theory. In A. Anastasi (Ed.), *Testing problems in perspective.* Washington, D.C.: American Council on Education, 1966.

Cronbach, L.J., & Glester, G.C. *Psychological tests and personnel decisions.* Urbana: University of Illinois Press, 1965.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles* New York: John Wiley & Sons, Inc., 1972.

DeWitt, L.J., & Weiss, D.J. A computer software system for adaptive ability measurement. *Research Report, 74-1,* Psychometric Methods Program, University of Minnesota, 1974.

Elwood, D.L. Automation of psychological testing. *American Psychologist,* 1969, 24, 287–289.

Feldt, L.S. The approximate sampling distribution of Kuder-Richardson Reliability Coefficient Testing. *Psychometrika,* 1965, 30, #3, 357–370.

Feldt, L.S. A note on the use of confidence bands to evaluate the reliability of a difference between two scores. *American Educational Research Journal,* 1967, 4, #2, 139–145.

Feldt, L.S. A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. *Psychometrika,* 1969, 34, #3, 363–373.

Ferguson, G.A. *Statistical analysis in psychology & education: 3rd edition.* New York: McGraw-Hill, 1971.

Ferguson, R.L. The development, implementation, and evaluation of a computer-assisted branched test for a program of individually prescribed instruction. (Doctoral dissertation, University of Pittsburgh), Ann Arbor, Mich.: University Microfilms, 1969, No. 70-4530.

Ferguson, R.L. Computer assistance for individualizing measurement. Report 1971/8, University of Pittsburgh, Learning and Research Development Center, March 1971.

Frary, R.B., & Zimmerman, D.W. Effect of variation in probability of guessing correctly on reliability of multiple-choice tests. *Educational and Psychological Measurements,* 1972, 9, 205–207.

Green, B.G., Jr. Comments on tailored testing. In W. H. Holtzman (Ed.), *Computer-assisted instruction, testing and guidance.* New York: Harper & Row, 1970.

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

Harman, H.H., Helm, C.E., & Loye, D.E. (Eds.), *Computer assisted testing.* Princeton, N.J.: Educational Testing Service, 1968.

Hick, W.E. Information theory and intelligence tests. *British Journal of Psychology, Statistical Section,* 1951, 4, 157–164.

Holtzman, W.H. Individually tailored testing: Discussion. In W. H. Holtzman (Ed.), *Computer-assisted instruction, testing, and guidance.* New York: Harper & Row, 1970.

Hoyt, C. Test reliability estimated by analysis of variance. *Psychometrika,* 1941, 6, #3, 153–160.

Hubbard, J.P. Programmed testing in the examinations of the National Board of Medical Examiners. In A. Anastasi (Ed.), *Testing problems in perspective.* Washington, D.C.: American Council in Education, 1966.

Kalisch, S.J. A tailored testing model employing the beta distribution. Unpublished paper. Educational Evaluation and Research Design Program, Florida State University, 1974.

Kappauf, W.E. Use of an on-line computer for psychological testing with the up-and-down method. *American Psychologist,* 1969, 24, 207–211.

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

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

Levine, R.D., & Lord, F.M. An index of the discriminating powers of a test at different parts of the score range. *Educational and Psychological Measurement*, 1959, **19**, 497–500.

Linn, R.L., Rock, D.A., & Cleary, T.A. *Sequential testing for dichotomous decisions.* College entrance examination board research and development report, RDR 69-70, No. 3, 1970 (ETS, RB-70-31).

Linn, R.L., Rock, D.A., & Cleary, T.A. Sequential testing for dichotomous decisions. *Educational and Psychological Measurement*, 1972, **32**, 85–96.

Lord, F.M. Do tests of the same length have the same standard errors of measurement? *Educational and Psychological Measurement*, 1957, **17**, 510–521.

Lord, F.M. Tests of the same length do have the same standard errors of measurement. *Educational and Psychological Measurement*, 1959, **19**, 233–239.

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

Lord, F.M. Robbins-Munro procedures for tailored testing. *Educational and Psychological Measurement.* 1971, **31**. (a)

Lord, F.M. The self-scoring flexilevel test. *Journal of Educational Measurement*, 1971, **8**, 147–151. (b)

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

Lord, F.M. A theoretical study of the measurement effectiveness of flexilevel tests. *Educational and Psychological Measurement*, 1971, **31**, 805–813. (d)

Lord, F.M. A theoretical study of two-stage testing. *Psychometrika*, 1971, **36**, 227–241. (e)

Lord, F.M. *Individualized testing and item characteristic curve theory.* Educational Testing Service, ETS-RB-72-50, Princeton, N.J., November, 1972.

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

McBride, J.R., & Weiss, D.J. A word knowledge item pool for adaptive ability measurement. *Research Report 74-2*, Psychometric Methods Program, University of Minnesota, 1974.

Moonan, W.J. Some empirical aspects of the sequential analysis technique as applied to an achievement examination. *Journal of Experimental Education*, 1950, **18**, 195-207.

Nie, N.H., Bent, D.H., & Hull, C.H. *Statistical package for the social sciences: SPSS.* New York: McGraw-Hill, 1970.

Nitko, A.J., & Feldt, L.S. A note on the effect of item difficulty distributions on the sampling distribution of KR-20. *American Educational Research Journal*, 1969, **6**, 433–437, May 1969.

Novick, M.R. *Bayesian Methods in Psychological Testing*, Princeton, N.J.: Educational Testing Testing, RB-69-31, 1969.

Nunnally, J.C. *Psychometric theory.* New York: McGraw-Hill, 1967.

Olivier, P. An overview of tailored Testing. Unpublished paper. Florida State University Program of Educational Evaluation and Research Design, July 1973.

Olivier, P. An evaluation of the self-scoring flexilevel tailored testing model. Unpublished doctoral dissertation, The Florida State University, 1974.

Owen, H.J. A Bayesian approach to tailored testing. Princeton, N.J.: *Educational Testing Service*, Research Bulletin, RB-69-72, 1969.

Owen, R.J. Bayesian sequential design and analysis of dichotomous experiments with special reference to mental testing. Unpublished paper, 1970.

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

Payne, W.H., & Anderson, D.E. Significance levels for the Kuder-Richardson twenty: an automated sampling empirical approach. *Educational and Psychological Measurement*, 1962, 28, 23-39.

SCAT Series II, *Cooperative school and college ability tests*, Princeton, N.J.: Educational Testing Service, 1967.

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

Siegel, S. Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill, 1956.

Stanley, J.C. Reliability. In R. I. Thorndike (Ed.), *Educational Measurement.* Washington, D.C.: American Council on Education, 1971.

Stanley, J.C., & Wang, M.D. Weighting test items and test item options: An overview of the analytical and empirical literature, *Educational and Psychological Measurement*, 1970, 30, 21–35.

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

Terman, L.M., & Merrill, M.A. *Stanford-Binet intelligence scale: Manual for the third revision form L-M.* Houghton-Mifflin, 1960.

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

Urry, V.W. *Individualized testing by Bayesian estimation.* Washington University, Seattle, Bureau of Testing Project: 0171–177, 3–29.

Wald, A. *Sequential analysis.* New York: Wiley, 1946.

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

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

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

Weiss, D.J., & Betz, N.E. Ability Measurement: Conventional or adaptive? *Research Report 73-1. Psycholometric Methods Program,* University of Minnesota, 1973.

Winer, B.J. *Statistical principles in experimental design:* 2nd Ed. New York: McGraw-Hill, 1971.

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

Wood, R. Computerized adaptive sequential testing. Unpublished doctoral dissertation, University of Chicago, 1971.

Wood, R. Response-contingent testing. *Review of educational research,* 1973, 43, #4, 529–544.

APPENDIX A: ITEM STATISTIC COMPARISON

ITEM NUM.	NORM GROUP		LINEAR GROUP			STRDPTV GROUP		
	N	P	N	P	S.E.	N	P	S.E.
01	3133	.86	11	1.00	0	***	***	***
02	3133	.86	10	1.00	0	***	***	***
03	3133	.92	10	1.00	0	***	***	***
04	3133	.93	10	1.00	0	***	***	***
05	3133	.92	10	.90	.10	1	1.00	0
06	3133	.81	10	1.00	0	***	***	***
07	3133	.79	10	1.00	0	***	***	***
08	3133	.70	10	.90	.10	***	***	***
09	3133	.85	10	1.00	0	***	***	***
10	3133	.83	10	.90	.10	***	***	***
11	3133	.79	10	1.00	0	15	.87	.10
12	3133	.71	10	.90	.10	***	***	***
13	3133	.73	10	.80	.13	***	***	***
14	3133	.77	10	.70	.15	***	***	***
15	3133	.75	10	.90	.10	***	***	***
16	3133	.68	10	.60	.16	***	***	***
17	3133	.56	10	.70	.15	40	.78	.07
18	3133	.72	10	.90	.10	***	***	***
19	3133	.58	10	.60	.16	1	1.00	0
20	3133	.64	10	1.00	0	3	.33	.33
21	3133	.58	10	.70	.15	1	1.00	0
22	3133	.58	10	.90	.10	2	0	0
23	3133	.60	10	1.00	0	1	0	0
24	3133	.58	10	.80	.13	***	***	***
25	3133	.63	10	.90	.10	1	1.00	0
26	3133	.70	10	.80	.13	***	***	***
27	3133	.58	10	.80	.13	***	***	***
28	3133	.60	10	1.00	0	3	1.00	0
29	3133	.68	10	1.00	0	***	***	***
30	3133	.48	10	.40	.16	4	1.00	0
31	3133	.62	10	.90	.10	***	***	***
32	3133	.52	10	.70	.15	39	.49	.08
33	3133	.53	10	.60	.16	***	***	***
34	3133	.51	10	.70	.15	***	***	***
35	3133	.46	10	.50	.17	54	.56	.07
36	3133	.38	10	.40	.16	43	.49	.08
37	3133	.55	10	.70	.15	***	***	***
38	3133	.42	10	.70	.15	37	.38	.08
39	3133	.40	10	.30	.15	2	0	0
40	3133	.52	10	.80	.13	32	.50	.09
41	3133	.35	10	.40	.16	34	.41	.09
42	3133	.53	10	.60	.16	***	***	***
43	3133	.45	10	.30	.15	14	.57	.14
44	3133	.38	10	.60	.16	40	.50	.08
45	3133	.40	10	.70	.15	54	.52	.07
46	3133	.35	10	.30	.15	21	.33	.11

ITEM NUM.	NORM GROUP		LINEAR GROUP			STRDPTV GROUP		
	N	P	N	P	S.E.	N	P	S.E.
47	3133	.49	10	.70	.15	2	.50	.50
48	----	---	---	---	---	---	---	---
49	3133	.34	10	.40	.16	32	.41	.09
50	3133	.28	---	---	---	19	.26	.10
51	3133	.90	8	1.00	0	***	***	***
52	3133	.86	8	.88	.13	***	***	***
53	3133	.88	8	1.00	0	***	***	***
54	3133	.77	8	.75	.16	***	***	***
55	3133	.87	8	1.00	0	***	***	***
56	3133	.84	8	.75	.16	***	***	***
57	3133	.88	8	1.00	0	***	***	***
58	3133	.86	8	1.00	0	***	***	***
59	3133	.69	8	1.00	0	40	.93	.04
60	3133	.64	8	.88	.13	3	.67	.33
61	3133	.76	8	.75	.16	***	***	***
62	3133	.71	8	.88	.13	***	***	***
63	3133	.69	8	1.00	0	***	***	***
64	3133	.70	8	.38	.18	***	***	***
65	3133	.71	8	.75	.16	16	.63	.13
66	3133	.83	8	.75	.16	***	***	***
67	3133	.71	8	.63	.18	***	***	***
68	3133	.75	8	.75	.16	***	***	***
69	3133	.63	8	.88	.13	***	***	***
70	3133	.74	8	.75	.16	***	***	***
71	3133	.64	8	1.00	0	14	.57	.14
72	3133	.62	8	.88	.13	***	***	***
73	3133	.57	8	1.00	0	33	.70	.08
74	3133	.75	8	.88	.13	***	***	***
75	3133	.36	8	.38	.18	21	.52	.11
76	3133	.55	8	.25	.16	***	***	***
77	3133	.43	8	.63	.18	11	.55	.16
78	3133	.48	8	1.00	0	47	.64	.07
79	3133	.64	8	.63	.18	10	.60	.16
80	3133	.47	8	.88	.13	51	.69	.07
81	3133	.51	8	.50	.19	38	.66	.08
82	3133	.52	8	.63	.18	7	.14	.14
83	3133	.54	8	.63	.18	21	.67	.11
84	3133	.47	8	.25	.16	47	.53	.07
85	3133	.57	8	.25	.16	***	***	***
86	3133	.41	8	.38	.18	1	1.00	0
87	3133	.38	8	.63	.18	53	.49	.07
88	3133	.46	8	.50	.19	2	0	0
89	3133	.52	8	.75	.16	***	***	***
90	3133	.50	8	.63	.18	55	.67	.06
91	3133	.39	8	.63	.18	43	.56	.08
92	3133	.45	8	.38	.18	54	.43	.07

ITEM NUM.	NORM GROUP		LINEAR GROUP			STRDPTV GROUP		
	N	P	N	P	S.E.	N	P	S.E.
93	3133	.34	8	.25	.16	37	.30	.08
94	3133	.36	8	.63	.18	28	.46	.10
95	3133	.41	8	.50	.19	52	.67	.07
96	3133	.26	8	.25	.16	52	.39	.07
97	3133	.24	8	.25	.16	32	.34	.09
98	3133	.35	8	.75	.16	20	.40	.11
99	3133	.33	8	.75	.16	16	.31	.12
100	3133	.14	8	.38	.18	48	.42	.07
101	3133	.95	---	---	---	1	0	0
102	3133	.90	7	1.00	0	***	***	***
103	3133	.67	7	.88	.14	***	***	***
104	3133	.85	7	.71	.18	1	0	0
105	3133	.86	7	1.00	0	***	***	***
106	3133	.89	7	.57	.20	***	***	***
107	3133	.86	7	1.00	0	***	***	***
108	3133	.81	7	.86	.14	***	***	***
109	3133	.84	7	.57	.20	***	***	***
110	3133	.22	7	.43	.20	27	.67	.09
111	3133	.74	7	.88	.14	***	***	***
112	3133	.52	7	1.00	0	***	***	***
113	3133	.51	7	.57	.20	49	.61	.07
114	3133	.72	7	.57	.20	***	***	***
115	3133	.77	7	.86	.14	***	***	***
116	3133	.67	7	.71	.18	***	***	***
117	3133	.69	7	.86	.14	12	.50	.15
118	3133	.66	7	.29	.18	***	***	***
119	3133	.68	7	.57	.20	8	.63	.18
120	3133	.62	7	1.00	0	***	***	***
121	3133	.66	7	.43	.20	***	***	***
122	3133	.61	7	.29	.18	***	***	***
123	3133	.61	7	.57	.20	2	0	0
124	3133	.60	7	.43	.20	***	***	***
125	3133	.51	7	.43	.20	***	***	***
126	3133	.59	7	.43	.20	42	.26	.07
127	3133	.54	7	.29	.18	26	.54	.01
128	3133	.57	7	.29	.18	***	***	***
129	3133	.43	7	.14	.14	49	.53	.07
130	3133	.50	7	.14	.14	54	.50	.07
131	3133	.56	7	.29	.18	3	0	0
132	3133	.42	7	.29	.18	48	.38	.07
133	3133	.50	7	.57	.20	18	.33	.11
134	3133	.43	7	.29	.18	49	.14	.05
135	3133	.48	7	.43	.20	53	.55	.07
136	3133	.51	7	.86	.14	18	.78	.10
137	3133	.57	7	.86	.14	9	.56	.18
138	3133	.56	7	.71	.18	38	.55	.08

ITEM NUM.	NORM GROUP		LINEAR GROUP			STRDPTV GROUP		
	N	P	N	P	S.E.	N	P	S.E.
139	3133	.39	7	.57	.20	25	.32	.10
140	3133	.43	7	.14	.14	1	0	0
141	3133	.35	7	.86	.14	22	.68	.10
142	3133	.39	7	.29	.18	50	.36	.07
143	3133	.48	7	0	0	40	.23	.07
144	3133	.36	7	.71	.18	15	.67	.13
145	3133	.35	7	.57	.20	17	.41	.12
146	3133	.38	7	.29	.18	24	.54	.10
147	3133	.40	7	.86	.14	37	.78	.07
148	3133	.30	7	0	0	51	.14	.05
149	3133	.20	7	.14	.14	43	.40	.08
150	3133	---	---	---	---	---	---	---
151	3133	.89	13	.92	.08	***	***	***
152	3133	.80	13	.85	.10	***	***	***
153	3133	.89	13	.92	.08	***	***	***
154	3133	.85	13	1.00	0	***	***	***
155	3133	.84	13	.92	.08	5	.80	.20
156	3133	.88	13	1.00	0	***	***	***
157	3133	.77	13	1.00	0	***	***	***
158	3133	.83	13	.85	.10	***	***	***
159	3133	.82	13	1.00	0	***	***	***
160	3133	.83	13	.92	.08	***	***	***
161	3133	.77	13	.85	.10	***	***	***
162	3133	.79	13	1.00	0	2	1.00	0
163	3133	.70	13	.77	.12	***	***	***
164	3133	.80	13	1.00	0	***	***	***
165	3133	.65	13	.77	.12	1	1.00	0
166	3133	.74	13	.92	.08	***	***	***
167	3133	.75	13	.69	.13	***	***	***
168	3133	.70	13	.85	.10	9	.78	.15
169	3133	---	---	---	---	---	---	---
170	3133	.72	13	1.00	0	***	***	***
171	3133	.64	13	.92	.08	***	***	***
172	3133	.65	13	.92	.08	***	***	***
173	3133	.64	13	.69	.13	***	***	***
174	3133	.61	13	.54	.14	***	***	***
175	3133	.66	13	.92	.08	1	1.00	0
176	3133	.65	13	.77	.12	2	1.00	0
177	3133	.64	13	.92	.08	***	***	***
178	3133	.59	13	.69	.13	3	1.00	0
179	3133	.60	13	1.00	0	***	***	***
180	3133	.52	13	.39	.14	27	.67	.09
181	3133	.55	13	.92	.08	1	0	0
182	3133	.50	13	.85	.10	5	.80	.20
183	3133	.51	13	.69	.13	3	.33	.33
184	3133	.49	13	.69	.13	***	***	***

ITEM NUM.	NORM GROUP		LINEAR GROUP			STRDPTV GROUP		
	N	P	N	P	S.E.	N	P	S.E.
185	3133	.52	13	.92	.08	29	.66	.09
186	3133	.46	13	.39	.14	28	.71	.09
187	3133	.50	13	.85	.10	49	.65	.07
188	3133	.68	13	.85	.10	2	1.00	0
189	3133	.52	13	.62	.14	48	.69	.07
190	3133	.53	13	.85	.10	***	***	***
191	3133	.43	13	.92	.08	55	.78	.06
192	3133	.50	13	.77	.12[a]	***	***	***
193	3133	.41	13	.54	.14	32	.47	.10
194	3133	.53	13	.92	.08	***	***	***
195	3133	.33	13	.77	.12	10	.80	.13
196	3133	.30	13	.62	.14	55	.40	.07
197	3133	.34	13	.62	.14	12	.91	.08
198	3133	.24	13	.77	.12	23	.48	.11
199	3133	.25	13	.23	.12	52	.42	.07
200	3133	.22	13	.77	.12	30	.47	.09
201	3133	---	---	---	---	---	---	---
202	3133	.89	9	1.00	0	***	***	***
203	3133	.90	9	.89	.11	***	***	***
204	3133	.77	9	.67	.17	3	1.00	0
205	3133	.88	9	1.00	0	1	0	0
206	3133	.83	9	1.00	0	***	***	***
207	3133	.80	9	1.00	0	9	.89	.11
208	3133	.86	9	.78	.15	***	***	***
209	3133	.69	9	.78	.15	***	***	***
210	3133	.74	9	.67	.17	***	***	***
211	3133	.68	9	.78	.15	***	***	***
212	3133	.81	9	.78	.15	***	***	***
213	3133	.76	9	1.00	0	***	***	***
214	3133	.69	9	.67	.17	***	***	***
215	3133	.67	9	.89	.11	20	.90	.07
216	3133	.82	9	.89	.11	3	1.00	0
217	3133	.71	9	.89	.11	***	***	***
218	3133	.89	9	.89	.11	***	***	***
219	3133	.78	9	.67	.17	***	***	***
220	3133	.83	9	1.00	0	***	***	***
221	3133	.73	9	.89	.11	***	***	***
222	3133	.78	9	1.00	0	***	***	***
223	3133	.74	9	.78	.15	***	***	***
224	3133	.71	9	1.00	0	2	.50	.50
225	3133	.80	9	1.00	0	7	1.00	0
226	3133	.66	9	.89	.11	5	.80	.20
227	3133	.61	9	.78	.15	***	***	***
228	3133	.70	9	1.00	0	***	***	***
229	3133	.61	9	.44	.18	***	***	***
230	3133	.65	9	.67	.17	13	.62	.14

ITEM NUM.	NORM GROUP		LINEAR GROUP			STRDPTV GROUP		
	N	P	N	P	S.E.	N	P	S.E.
231	3133	.58	9	.67	.17	9	.89	.11
232	3133	.52	9	.78	.15	16	.50	.13
233	3133	.41	9	.56	.18	18	.50	.12
234	3133	.57	9	.78	.15	5	.60	.25
235	3133	.48	9	.78	.15	43	.86	.05
236	3133	.58	9	.89	.11	***	***	***
237	3133	.54	9	.78	.15	23	.91	.06
238	3133	.43	9	.78	.15	4	1.00	0
239	3133	.52	9	.78	.15	15	.53	.13
240	3133	.42	9	.67	.17	7	.86	.14
241	3133	.61	9	.67	.17	28	.82	.07
242	3133	.44	9	.67	.17	23	.57	.11
243	3133	.34	9	.67	.17	37	.70	.08
244	3133	.28	9	.56	.18	53	.55	.07
245	3133	.45	9	.67	.17	53	.72	.06
246	3133	.37	9	.33	.17	11	.73	.14
247	3133	.40	9	.56	.18	33	.58	.09
248	3133	.42	9	.56	.18	1	0	0
249	3133	.23	---	---	---	17	.35	.12
250	3133	---	---	---	---	---	---	---

*** This item not presented to stradaptive subject.
--- This item removed from stradaptive pool.

APPENDIX B: TRANSFORMATION OF TRADITIONAL ITEM DIFFICULTY (P_g) AND BISERIAL CORRELATION (r'_g) TO NORMAL OGIVE PARAMETERS b_g AND a_g

Item Number	P_g	b_g	r'_g	a_g
1	.82	-1.50	.61	.77
2	.82	-1.39	.66	.88
3	.88	-2.50	.47	.53
4	.89	-2.27	.54	.64
5	.88	-1.70	.69	.95
6	.77	-1.91	.62	.79
7	.75	- .95	.71	1.01
8	.66	- .86	.48	.55
9	.81	-1.91	.46	.52
10	.79	-1.55	.52	.61
11	.75	-1.05	.64	.83
12	.67	-.75	.59	.73
13	.69	-1.18	.42	.46
14	.73	-1.09	.56	.68
15	.71	- .88	.63	.81
16	.64	- .64	.56	.68
17	.52	- .10	.52	.61
18	.68	- .73	.64	.83
19	.54	- .18	.56	.68
20	.60	- .55	.46	.52
21	.54	- .31	.32	.34
22	.54	- .16	.64	.83
23	.56	- .30	.50	.58
24	.54	- .23	.44	.49
25	.59	- .54	.42	.46
26	.66	- .83	.50	.58
27	.54	- .16	.63	.81
28	.56	- .24	.62	.79
29	.64	- .85	.42	.46
30	.44	.32	.47	.53
31	.58	- .35	.57	.69
32	.48	.09	.58	.71
33	.49	.05	.54	.64
34	.47	.12	.61	.77
35	.42	.43	.47	.53
36	.34	.88	.47	.53
37	.51	- .04	.66	.88
38	.38	.62	.49	.56
39	.36	.66	.54	.64
40	.48	.11	.46	.52

Item Number	P_g	b_g	r'_g	a_g
41	.31	1.60	.31	.33
42	.49	.06	.45	.50
43	.41	.95	.24	.25
44	.34	.79	.52	.61
45	.36	.65	.55	.66
46	.31	1.27	.39	.42
47	.45	.29	.43	.48
48	.14	4.71	.23	.24
49	.30	1.34	.39	.42
50	.24	2.94	.24	.25
51	.86	-1.69	.64	.83
52	.82	-1.50	.61	.77
53	.84	-1.51	.66	.88
54	.73	-1.02	.60	.75
55	.83	-1.36	.70	.98
56	.80	-1.40	.60	.75
57	.84	-1.78	.56	.68
58	.82	-1.79	.51	.59
59	.65	- .62	.62	.79
60	.60	- .42	.60	.75
61	.72	-1.17	.50	.58
62	.67	- .75	.59	.73
63	.65	- .62	.62	.79
64	.66	-1.06	.39	.42
65	.67	- .90	.49	.56
66	.79	-2.07	.39	.42
67	.67	- .60	.73	1.07
68	.71	- .91	.61	.97
69	.59	- .37	.62	.79
70	.70	- .95	.55	.66
71	.60	- .44	.57	.69
72	.58	- .53	.38	.41
73	.53	- .17	.45	.50
74	.71	-1.35	.41	.45
75	.32	.75	.62	.79
76	.51	- .06	.45	.50
77	.39	.67	.42	.46
78	.44	.28	.54	.64
79	.60	- .49	.52	.61
80	.43	.30	.58	.71
81	.47	.12	.63	.81
82	.48	.13	.40	.44
83	.50	0.00	.38	.41
84	.43	.38	.46	.52
85	.53	- .18	.43	.48

Item Number	P_g	b_g	r'_g	a_g
86	.37	.56	.59	.73
87	.34	.76	.54	.64
88	.42	.37	.55	.66
89	.48	.10	.48	.55
90	.46	.20	.51	.59
91	.35	.88	.44	.49
92	.41	.63	.36	.39
93	.33	1.57	.28	.29
94	.32	1.06	.44	.49
95	.37	.81	.41	.45
96	.22	1.76	.44	.49
97	.20	1.83	.46	.52
98	.31	1.91	.26	.27
99	.29	2.13	.26	.27
100	.10	2.91	.44	.49
101	.91	-5.16	.26	.27
102	.86	-3.28	.33	.35
103	.63	- .77	.43	.48
104	.81	-2.31	.38	.41
105	.82	-2.03	.45	.50
106	.85	-3.57	.29	.30
107	.82	-2.12	.43	.48
108	.77	-2.17	.34	.36
109	.80	-1.22	.69	.95
110	.18	2.29	.40	.44
111	.70	-1.31	.40	.44
112	.48	.13	.40	.44
113	.47	.24	.32	.34
114	.68	-1.38	.34	.36
115	.73	-1.80	.34	.36
116	.63	- .61	.54	.64
117	.65	- .88	.44	.49
118	.62	- .60	.51	.59
119	.64	- .80	.45	.50
120	.58	- .65	.31	.33
121	.62	- .90	.34	.36
122	.57	- .33	.54	.64
123	.57	- .57	.31	.33
124	.56	- .49	.31	.33
125	.47	.14	.55	.66
126	.55	- .33	.38	.41
127	.50	0.00	.36	.39
128	.53	- .19	.40	.44
129	.39	.87	.32	.34
130	.46	.31	.32	.34

Item Number	P_g	b_g	r_g	a_g
131	.52	- .10	.48	.55
132	.38	.71	.43	.48
133	.46	.21	.47	.53
134	.39	.59	.47	.53
135	.44	.34	.44	.49
136	.47	.13	.56	.68
137	.53	- .21	.36	.39
138	.52	- .11	.45	.50
139	.35	1.24	.31	.33
140	.39	.50	.56	.68
141	.31	1.01	.49	.56
142	.35	1.13	.34	.36
143	.44	.29	.52	.61
144	.32	1.34	.35	.37
145	.31	1.03	.48	.55
146	.34	1.33	.31	.33
147	.36	1.23	.29	.30
148	.26	1.61	.40	.44
149	.16	3.68	.27	.28
150	INFORMATION NOT AVAILABLE -- NOT USED			
151	.89	-2.31	.53	.63
152	.80	-1.56	.54	.64
153	.89	-1.83	.67	.90
154	.85	-1.52	.68	.93
155	.84	-1.74	.57	.69
156	.88	-2.03	.58	.71
157	.77	-1.27	.58	.71
158	.83	-1.87	.51	.59
159	.82	-2.03	.45	.50
160	.83	-1.65	.58	.71
161	.77	-1.30	.57	.69
162	.79	-1.97	.41	.45
163	.70	- .90	.58	.71
164	.80	-1.68	.50	.58
165	.65	- .76	.51	.59
166	.74	-1.46	.44	.49
167	.75	-1.65	.41	.45
168	.70	-1.34	.39	.42
169	DISCRIMINATION INDEX TOO LOW -- NOT USED			
170	.72	-1.04	.56	.68
171	.64	- .65	.55	.66
172	.65	- .74	.52	.61
173	.64	- .59	.61	.77
174	.61	-1.07	.26	.27
175	.66	- .92	.45	.50

Item Number	P_g	b_g	r_g^*	a_g
176	.65	- .80	.48	.55
177	.64	- .54	.66	.88
178	.59	- .41	.55	.66
179	.60	- .51	.50	.58
180	.52	- .16	.31	.33
181	.55	- .31	.41	.45
182	.50	0.00	.58	.71
183	.51	- .05	.49	.56
184	.49	.07	.34	.36
185	.52	- .11	.47	.53
186	.46	.29	.35	.37
187	.50	0.00	.40	.44
188	.68	- .84	.56	.28
189	.52	- .19	.27	.28
190	.53	- .14	.55	.66
191	.43	.38	.47	.53
192	.50	0.00	.41	.45
193	.41	.91	.25	.26
194	.53	- .18	.43	.48
195	.33	.94	.47	.53
196	.30	1.25	.42	.46
197	.34	.75	.55	.66
198	.24	1.91	.37	.40
199	.25	1.69	.40	.44
200	.22	1.27	.61	.77
201	.90	-4.13	.31	.33
202	.89	-1.95	.63	.81
203	.90	-1.97	.65	.86
204	.77	-1.39	.53	.63
205	.88	-1.90	.62	.79
206	.83	-1.47	.65	.86
207	.80	-1.87	.45	.50
208	.86	-1.90	.57	.69
209	.69	-1.20	.41	.45
210	.74	-1.37	.47	.53
211	.68	-1.95	.24	.25
212	.81	-1.83	.48	.55
213	.76	-1.31	.54	.64
214	.69	-1.08	.46	.52
215	.67	- .72	.61	.77
216	.82	-2.08	.44	.49
217	.71	- .91	.61	.77
218	.89	-2.36	.52	.61
219	.78	-2.03	.38	.41
220	.83	-1.95	.49	.56

Item Number	P_g	b_g	r'_g	a_g
221	.73	- .97	.63	.81
222	.78	-1.46	.53	.63
223	.74	-1.69	.38	.41
224	.71	- .86	.64	.83
225	.80	-1.11	.76	1.17
226	.66	- .69	.60	.75
227	.61	- .76	.37	.40
228	.70	- .83	.63	.81
229	.61	- .53	.53	.73
230	.65	- .65	.59	.73
231	.58	- .42	.48	.55
232	.52	- .11	.46	.52
233	.41	.56	.41	.45
234	.57	- .35	.51	.59
235	.48	.09	.57	.69
236	.58	- .44	.46	.52
237	.54	- .16	.64	.83
238	.43	.30	.59	.73
239	.52	- .08	.61	.77
240	.42	.40	.51	.59
241	.61	- .49	.57	.69
242	.44	.27	.56	.68
243	.34	.71	.58	.71
244	.28	1.19	.49	.56
245	.45	.25	.51	.59
246	.37	.79	.42	.46
247	.40	.53	.48	.55
248	.43	.56	.36	.39
249	.23	1.94	.38	.41
250	.14	4.71	.23	.24

APPENDIX C: .FORM LETTER

July 28, 1974

Dear Orientation Participant,

This note is a request for your help. I am a graduate student at FSU, working on a research project. I desperately need participants to volunteer to help me.

If you are willing to help, I will need from 30 to 45 minutes of your time sometime during the three-day orientation program. You will operate an electronic computer terminal for this study. The experience should be interesting and informative for you, and may simplify your computer usage while a student here at Florida State.

If you are interested in learning more about this project, please meet with me at Moore Auditorium (in the Union Complex) at 9:30 A. M. on Monday, the 29th. I will explain all about the project and answer any questions that you may have. Thanks again.

Brian Waters

APPENDIX D: DESCRIPTION OF DATA

Description of Data Stored on each Testee's Data File

Data is stored in 10-character words

Word No.	Data
1	Identification number or Social Security number.
2	Keyword as entered by proctor.
3	Current location in program: 0-1000 instructions 1001-2000 Test 1 2001-3000 Test 2, if given 3001-4000 Post-Feedback
4	Elapsed time in seconds from time subject began instructions until testing was completed.
5	Total time in seconds spent on instructions.
6	Total time in seconds spent on test 1.
7	Number of errors on instructional screens 1-10. 1 character per screen.
8	Number of errors on instructional screens 11-20. 1 character per screen.
9	Number of items correct on test 1.
10-12	Testee's name, 30 characters.
13	Characters 1-2: subject's estimated ability, if taken. Characters 3-8: blank Characters 9-10: college code (01-27)
14	Social security number, if available.
15	Date of testing
16	Seconds since midnight when testing began.
17	Elapsed time in seconds spent on test 2.
18	Maximum number of questions which could be given on test 1.
19	Maximum number of questions which could be given on test 2.
20	Number of items attempted on test 1.
21	Number of items attempted on test 2.
22	First score on test 1.
23	First score on test 2.
24	(reserved for program for recovery information)
25	Number of items correct on test 2.
26	Second score on test 1.
27	Second score on test 2.
28-30	(reserved for program for recovery information)

Data on each vocabulary item is packed into one word as follows:

character 1: response code

code	meaning
0	Item answered incorrectly
1	Item answered correctly
2	Item answered with a ?.

2: actual response (1-5, ?+0)
3-6: reference number of item presented
7: number of presentations of screen
8-10: response latency in seconds

☆U.S. GOVERNMENT PRINTING OFFICE: 1975-671-602/8031