# ABILITY MEASUREMENT: CONVENTIONAL OR ADAPTIVE?

David J. Weiss

and

Nancy E. Betz

Research Report 73-1

Psychometric Methods Program
Department of Psychology
University of Minnesota
February 1973

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

## Abstract

Research to date on adaptive (sequential, branched, individualized, tailored, programmed, response-contingent) ability testing is reviewed and summarized, following a brief review of problems inherent in conventional individual and group approaches to ability measurement. Research reviewed includes empirical, simulation and theoretical studies of adaptive testing strategies. Adaptive strategies identified in the literature include two-stage testing, and multistage tests. Multistage tests are differentiated into fixed branching models and variable branching models (including Bayesian and non-Bayesian strategies). Results of research using the various strategies and research approaches are compared and summarized, leading to the general conclusion that adaptive testing can considerably reduce testing time and at the same time yield scores of higher reliability and validity than conventional tests, under a number of circumstances. A number of new psychometric problems raised by adaptive testing are discussed, as is the criterion problem in evaluating the utility of adaptive testing. Problems of implementing adaptive testing in a paper and pencil mode, or using special testing machines are reviewed; the advantages of computer-controlled adaptive test administration are described.

# Contents

## Ability Measurement: Conventional or Adaptive?

Ability measurement began with the work of Binet, who developed the first scale that correlated importantly with the criteria considered to indicate intellectual or scholastic ability. Binet's scale and its revisions are administered to one individual at a time within a procedural framework that requires the examiner to adapt his administration to the characteristics of the individual being tested. Thus, they may be thought of as "adaptive" individual tests.

The 1960 revision of the Stanford-Binet (Terman & Merrill, 1960) provides an example of the adaptive individual test. First, the level at which to begin testing varies according to the administrator's judgment of the testee's ability; the idea is to begin at a level where the testee is relatively likely to succeed. Second, the order of item presentation is not fixed but depends to some extent on the testee's performance on and reaction to previous items. The extent of item presentation is controlled by a determination of basal and ceiling ages such that few items are presented at levels which are either much too hard or much too easy for the individual in question. Further, the administrator is often able to maintain or increase the subject's motivation for doing well by providing encouragement and feedback when necessary. Finally, there are no set limits on testing time (although a few subtests do have a time limit), but response times are frequently recorded as part of the psychometric data obtained.

Other individual tests followed the Binet, but most of these retained only some of the features of its approach. The Wechsler Adult Intelligence Scale (Wechsler, 1955), for example, is less adaptive and more standardized. The starting point for each subtest is fixed, although some subtests contain a provision to administer 3 or 4 very easy items if the first 1 or 2 regular items are failed. There is no flexibility in item or subtest order except in the above instance, i.e., normally each person is administered the same sequence of items. Neither is there a determination of basal and ceiling ages, although in most subtests a certain number of consecutive failures constitutes a basis for stopping that subtest. It is likely, then, that many subjects take a large number of items that are far too easy for them, and that some subjects may be tested beyond their ability level by one chance success amidst a string of failures. The Wechsler scales, like the Binet, provide for encouragement of the subject and for measurement of response times within an untimed test, but the use of a more standard administration procedure makes it appropriate to think of them as "standardized" or "conventional" individual tests.

There are several problems inherent in individually administered tests, whether using "adaptive" or "standardized" administration. Probably most obvious is the fact that they must be administered by a highly-trained examiner to one person at a time, and this is both expensive and inefficient. Group tests, which permit efficient mass administration by examiners who need only a minimum of training, were developed primarily to solve this problem. However, a variety of other more subtle problems with individual tests can be ascribed to examiner variables, which introduce error variance into the determination of ability level.

Problems in Individual Tests

There is evidence that different examiners score items on individual tests in different ways, even though they are following the same instructions. Studies in which such examiner differences are reported include those by Cieutat (1965), Cohen (1950), Plumb and Charles (1955), Schwartz (1966), Smith and May (1967), and Walker, Hunt & Schwartz (1965). Only two studies (Murdy, 1962; Nichols, 1959) did not find significant differences in examiner scoring on individual tests. Some of these examiner scoring differences may be due to an expectancy effect (Sattler, Hillix, & Neher, 1970; Sattler & Winget, 1970; Simon, 1969), or to knowledge of the testee's past performance (Egeland, 1969). Some studies also suggest that scoring might be biased by the examiner's feelings toward his subjects (Donahue & Sattler, 1971; Masling, 1959). In general, the data suggest that different examiners use different scoring strategies, and that these differences are sometimes complicated by examiner susceptibility to expectancies or personal feelings.

At least in the testing of children, the degree of rapport between tester and testee can influence the results of individual ability testing (Exner, 1966; Hata, Tsudzuki, Kuze, & Emi, 1958; Sachs, 1952; Tsudzuki, Hata, & Kuze, 1956), although an early study by Marine (1929) failed to show rapport effects. Similarly the test administrator's "adjustment" can affect test scores (Young, 1959), as can tester-testee sex differences or similarities (Quereshi, 1968; Stevenson & Allen, 1964). The data on examiner race yield conflicting results with several studies reporting no race effects (Caldwell & Knight, 1970; Miller & Phillips, 1966), and others reporting significant differences in test scores when testers were of different races (Forrester & Klaus, 1964; LaCrosse, 1964; Sattler, 1966). That examiner race has effects on ability test scores in interaction with situational stress is suggested by results reported by Katz & Greenbaum (1963) and Katz, Roberts, & Robinson (1965).

The available evidence suggests, then, that the score an individual receives on an individually-administered test may in some cases be heavily dependent on variables associated with the examiner or with the relationship between the testee and the examiner. Testing is usually a stressful situation which might intensify the tendency of examiner variables to introduce unsystematic and unwanted variance into the measurement of ability using individual tests.

## Limitations of Group Testing

Some psychologists realized before World War I that there was a need for a mode of testing more efficient and less expensive than individual tests. When the war began, however, the pressing need for rapid classification of the 1.5 million American recruits made group testing an imperative. An unpublished group test by Arthur S. Otis was the prototype for the development of the Army Alpha and Army Beta. These tests appeared about 1918 and started a period of tremendous growth in both the number and quality of group tests (Dubois, 1970).

Group tests are characterized by their high degree of standardization. They are administered to large numbers of people simultaneously by an examiner who attempts to follow an explicit set of examination procedures. The characteristics of group test procedures usually include: 1) a fixed set of items in a fixed order, 2) paper and pencil administration with separate answer sheets, 3) a fixed and "fair" set of time limits, and 4) completely objective scoring, usually done by machine, due to the popular multiple-choice item format.

Group tests control some of the variables present in individual tests, i.e., scoring, time limits, and number and order of items, but they too have a number of problems which frequently operate to differentially increase error.

Administrator variables. Although group tests were supposed to eliminate examiner effects, there is still some possibility that administrators can affect test scores through sex or race differences or through differences in the tendency of examiners to inadvertently arouse anxiety in susceptible individuals. There has been very little research in this area, but two studies have relevance for the problem. Baratz (1967) found that Negroes given the Test Anxiety Questionnaire (Mandler & Sarason, 1952) reported significantly greater anxiety when the examiner was white than when he was Negro. And Katz, Robinson, Epps, & Waly (1968) gave a hostility test disguised as a concept formation test to Negro high school students. With neutral instructions, the examiner's race had no effect. But under intelligence test instructions, significantly more hostility was shown with a Negro examiner than with a white examiner. These results suggested to the authors that Negro students inhibit hostile feelings in the presence of whites, and it is possible that the emotional conflict involved in controlling hostility may have disruptive effects on test performance.

These results are only suggestive, and there is a need for more research in this area.

Answer sheet effects. Different types of answer sheets have effects on standardized test performance, particularly with some groups of people. Gordon (1958) tested right-handed and left-handed naval recruits on the speeded Clerical test of the Navy's Basic Test Battery. The standard answer sheet for this test is a right-handed insert-type. The left-handed subjects performed significantly less well on the Clerical Test but just as well as the right-handed subjects on tests which did not use right-handed answer sheets. Merwin (1967) found significant differences between IBM 805 and MRC answer sheets on the Clerical Speed & Accuracy subtest of the Differential Aptitude Test (DAT) but found no differences on several unspeeded DAT subtests. Hayward (1967) administered an unspeeded reading test standardized on the IBM 805 answer sheet using IBM 805, IBM 1230, and Digitek answer sheets. She found that answer sheet and answer sheet by sex interaction effects were significant. Clark (1968) found that children with IQ's between 70 and 100 did significantly better when they could mark their answers on the test booklet as opposed to using a separate answer sheet. Whitcomb (1958) found that one group of adult males taking readings from a clock took an average of 120 seconds to record their answers on an IBM answer sheet. Another group, who wrote their responses in longhand, took only 60 seconds. Whitcomb concludes that when an IBM answer sheet is used with certain speeded tests, one may be measuring primarily answer sheet marking ability. Two groups of college students, a group of high school students, and a group of teenagers classified as "mentally retarded" were given 4 subtests of the GATB in a study by Nitardy, Peterson, & Weiss (1969). Separate answer sheets were eliminated for half of each group, and it was found that the groups were differentially affected on different subtests by this modification.

Item arrangement. The selection and sequencing of test items can affect test scores, both for the group as a whole and for certain individuals within the group.

Several studies have shown that different item arrangements affect the level of group performance on a single test. Sax & Cromach (1966) found that performance on the Henmon-Nelson Tests of Mental Ability under timed conditions was significantly better when items were arranged in ascending order of difficulty than when they were arranged in descending order of difficulty. Under very generous time limits, however, there were no significant differences. MacNichol (1956) found that under nearly pure power conditions, a hard-to-easy arrangement was significantly more difficult than an easy-to-hard arrangement.

Flaugher, Melton, & Myers (1968) rearranged SAT Verbal items so that items occurred in random order within blocks rearranged from the standard block order. A person taking the test with items arranged into 5 item blocks of like type in ascending order of difficulty (standard order) would have a 5.6 point advantage over a person of equal ability taking the test in the rearranged order. Since the SAT is speeded, the authors conclude that subjects may fail to reach items of different difficulties, and that this will affect their scores.

In a study by Sax & Carr (1962), college students attempted significantly more items and received significantly higher scores on the Henmon-Nelson Tests when items were arranged in a spiral omnibus format than when they were arranged in ascending order of difficulty within subtests. Sax & Carr offer the interpretation that as items get more difficult, the spiral omnibus format offers a variety of types of items; the student failing a number of difficult math items has more motivation to continue if he gets a vocabulary item instead of another math item. On the other hand, no differences in student performance on achievement tests were found when items were arranged in different orders in studies by Brenner (1964), Huck & Bowers (1972), and Smouse & Munz (1968).

Certain item arrangements may interact with particular types of testee characteristics. Peters & Messier (1970) found that students high on debilitating anxiety performed significantly less well than other students when items were arranged randomly but not when items were arranged sequentially. Results by Munz & Smouse (1968) indicated that students high on debilitating anxiety scored significantly lower on a final course examination than students high on facilitating anxiety when items were arranged randomly or from easy to hard but not when items were arranged from hard to easy. In contrast, however, are studies by Berger, Munz, Smouse, & Angelino (1969) and Marso (1970) in which different item orders did not differentially affect the performance of anxious and nonanxious students.

Klosner & Gellman (1971) hypothesized that item arrangement would be more important in the performance of low achievers than it is for high achievers. Their hypothesis was based on the proposal that item order has more effect under speeded conditions, and that for poorer students even a power test may seem speeded. They found that poorer students did best on ascending order of difficulty within subject matter order and worst when all items were arranged in order of difficulty. Item format made little difference with high achieving students.

In general, particular item arrangements can differentially affect group scores; this problem has relevance both for test construction and for the practice of administering alternate forms of a test for security purposes. More serious, though, is the possibility that certain item arrangements may have especially detrimental effects on people who are more susceptible to situational stress, i.e., the highly anxious or the poorly achieving.

Timing and time limits. Most group tests are timed solely for convenience of administration. Many psychometricians would probably agree that an untimed or "power" test is more appropriate for most abilities since most of the criteria to be predicted from ability tests are not heavily speeded. Time limits may penalize the slower but more accurate individual while benefiting the faster individual who has a tendency to guess. They may also penalize the person who tends to become anxious, and time limits can contribute to undesirable failure stress.

Some of the research in this area has been done with individual tests, but the findings would seem to apply in any testing situation which involves some degree of speededness.

Sarason, Mandler, & Craighill (1952) found that low anxiety subjects performed better on a digit-symbol substitution task when they were told that they were expected to finish within the given time limits, but high-anxiety subjects did better when told that they were not expected to finish. Siegman (1956) divided the WAIS into timed & untimed subtests and found that only high anxiety subjects had significantly lower scores on the timed subtests. He suggests that anxiety has a disruptive effect on performance on timed intelligence tests. Morris & Liebert (1969) administered the timed subtests of the WAIS to a group of subjects, only half of whom knew they were being timed. Subjects classified as "high worry" according to the Taylor Manifest Anxiety Scale did better when they did not know they were being timed, and "low worry" subjects did better when they did know. The worry by time interaction was even more pronounced when the tests were hard rather than easy.

Similar effects have been found for group tests. Matarazzo, Ulett, Guze, & Saslow (1954) found that level of anxiety was negatively correlated with scores on a timed scholastic aptitude test (ACE) but unrelated to scores on an untimed intelligence test (an abbreviated form of the Wechsler-Bellevue) or to college grade point average. Similarly, Sarason & Mandler (1952) found that low-anxious subjects did significantly better on the SAT, the Mathematical Aptitude Test, and the Henmon-Nelson Tests, all of which are speeded, than did high anxious subjects. However, there was no relationship between anxiety level and grades. Finally, students high on facilitating anxiety scored significantly higher on the timed Henmon-Nelson Tests than students high on debilitating anxiety in a study by Berger, Munz, Smouse, & Angelino (1969).

Standard set of items. Group tests usually require that the same set of items be given to all individuals in the group. But the standard set of items is typically aimed at the average individual in some specified population, and it is questionable whether these items are appropriate for individuals who deviate significantly from the average. Stanley (1971) suggests that the effective length of any test is considerably less than the total number of items for any given testee; he further asserts that administering all items to all testees is wasteful of both time and money.

Accuracy of measurement might also be affected by a standard set of items. Several reports (Baker, 1964; Levine & Lord, 1959; Lord, 1957, 1959, 1960) have concluded that the precision or reliability of measurement is not the same at all points on a score distribution, i.e., the standard error of measurement for a given individual is partially dependent on his "true" score. Thorndike (1951) and Davis (1952), among others, have shown that when item difficulties are concentrated at a given level, the standard error of measurement will be minimum for individuals at that point on the ability scale. On the typical "peaked" standard test, with item difficulties concentrated around .5, the error of measurement should be minimum for people of average ability and will increase as people deviate from the average. Ability estimates for subjects of high and low ability will consequently be less reliable than estimates for subjects of average ability.

When test items are too difficult for a given testee, the possibility of chance success through guessing on multiple-choice tests also contributes error differentially. Guessing reduces the reliability and validity of measurement for all subjects (Ebel, 1969; Frary & Zimmerman, 1970; Lord, 1957, 1963), but the increase in error is particularly pronounced for low ability subjects. According to Nunnally (1967), if all items are attempted, low ability subjects will guess the most because they know the least. Their scores will thus contain more error than those of average and high ability subjects. Boldt (1968) used formula scoring of multiple-choice items and found that error was greatest for people in the chance range, i.e., where a given score could have been obtained solely through random guessing. Thus, the use of a standard set of items for groups differing in ability can contribute to imprecise measurements.

## Summary

Group tests, then, have not provided a completely satisfactory solution to the problems in individual tests. Further, their high degree of standardization has introduced the problems of time limits, answer sheets, item set, and item arrangement as they affect whole groups and as they affect certain subgroups of individuals.

Adaptive individual testing, as represented by the Stanford-Binet, is still considered best by many because it is flexible enough to accommodate individual differences in ability and reaction to the testing process. But its subjectivity and susceptibility to examiner variables renders it unsatisfactory in terms of traditional psychometric criteria. Conventional individual tests, i.e., the Wechsler scales, retain the individual aspect and the recording of response times but lose much of the flexibility of the adaptive approach. Group tests sacrifice flexibility completely to gain convenience of administration and objectivity of scoring.

For several reasons, individual tests appear to be fairer than group tests, and in view of the current prevalence of criticism of psychological testing, fairness is a characteristic that must be considered. First, since individual tests are essentially untimed, their tendency to differentially arouse anxiety is probably much less than that of group tests. Second, group tests frequently have undesirable motivational effects when items are too hard or too easy for some individuals. Individual tests can maintain motivation at a more constant level by adapting item difficulty to subject ability. Further, group tests may be "off target" for some individuals in that there may be few or no items relevant to high and low ability subjects. Because of lower accuracy at the extremes, this may lead to highly unreliable measurement for those individuals. Some group tests are constructed with equal numbers of items at each ability level; this practice equalizes measurement accuracy but makes the test extremely long and wastes time that could be spent in more productive ways. With a good individualized test, testing time could be minimized without sacrificing accuracy by giving an individual only those items that are relevant to his ability. This would also decrease guessing considerably since people guess most when items are too difficult for them.

Neither conventional individual nor group tests appear to offer satisfactory alternatives to the adaptive approach; sacrificing flexibility for standardization seems to create as many problems as it solves. The potential advantages of the adaptive or individualized approach are clear. The problems that remain are to demonstrate the utility of the approach on traditional psychometric criteria, and to find a mode of implementation that minimizes or removes the subjectivity and examiner variance which have plagued individual testing.

## Background and Language of Adaptive Testing

Background. Adaptive testing involves varying test item presentation procedures according to characteristics of the individual being tested. In the majority of studies of adaptive testing test items are chosen for administration to a given testee based on that individual's responses to the previous item, or to a set of previous items. This approach builds on the basic logic implicit in Binet's work, in which the level of difficulty of succeeding questions is based on the testee's performance on blocks of previous test questions.

It is not surprising, therefore, that attempts to adapt ability tests to characteristics of the testee arose from clinical applications of individual ability tests. Spache (1942) was concerned with the effect of successive failure on scores on the Stanford-Binet. To determine whether successive failures might have an effect on Stanford-Binet scores, he re-scored test protocols as if 1 or 2 easy items had been presented whenever the testee failed 2 or 3 items in succession. His analysis showed no significant differences in test scores, but he concluded that the adaptive method was better since it would avoid large numbers of consecutive failures. Spache's study is limited, however, in that it did not involve actual adaptive administration; the study also used a group of gifted children, and it could be expected that adaptive testing might have greater effects with other groups.

Hutt (1947) actually administered Stanford-Binet items adaptively. When a child failed an item, he was given an easier one, and when he passed he was given a harder one. Testing was ended with items close to the subject's mental age, so that the end of the test would not be characterized by frustration and failure as is usually the case. Adaptive testing did not yield higher IQ's in a group of well-adjusted school children, but poorly adjusted children received reliably higher IQ's with the adaptive method.

A group of older people, aged 65-75, was studied by Greenwood & Taylor (1965) using an adaptive administration of the WAIS. The control group was given the standard WAIS initially and again after a 3-month interval. The experimental group was given the standard WAIS initially but an adaptive WAIS on the retest. In the adaptive WAIS each subtest was begun with an item below the testee's anticipated ability level; easy and hard items were then alternated, and a pool of nonscored easy items was available to ensure that the examiner would not run out of easy items. Retest scores of the adaptive group were significantly higher than those of the control group. The study was then repeated with college students, but no differences were found. This latter finding supports the possibility that Spache's (1942) inability to find differences in his simulated adaptive administration was due to the high ability characteristics of the group tested.

Frandsen, McCullough & Stone (1950) tried a serial administration of the Stanford-Binet in which all similar items were given together. This procedure avoids placing all of the most difficult items at the end of the test, as in the standard consecutive order. Although there were no significant differences between the results obtained from standard and serial administration for a group of normal children, the authors conclude that psychometrists can therefore continue to use the same norms while reducing the anxiety and frustration resulting from ending the test with a long succession of failures.

Outside the realm of clinical ability test administration, adaptive ability testing appears to have generated considerable interest, speculation and research. As early as 1951, Hick suggested that ability tests be redesigned in order to extract maximum "information" from a set of responses to ability test items. Based on findings in signal detection theory and information theory, Hick suggested that a testee be given a more difficult test item if he gets a previous item correct, or an easier item following an incorrect response. In constructing tests, he suggested that each test question have a .50 chance of being correctly answered by those who answered a previous test item correctly. Building on a different set of premises derived from decision theory, Cronbach (1966) suggested in 1954 that ability test administration could provide more information in a given unit of time if testing procedures were adapted to characteristics of the individual. Cronbach's suggestions included the design of a series of short screening tests to be administered within an hierarchical abilities model, followed by more intensive measurement tests for individuals who obtained high scores on specific screening tests.

Recent literature on adaptive testing includes a variety of kinds of studies as well as a variety of terms to refer to the concept of adaptive testing. Adapting ability test items to characteristics of the individual has been referred to as sequential testing (Krathwohl & Huyser, 1956; Paterson, 1962), branched testing (Bayroff, 1964), individualized measurement (Weiss, 1969), tailored testing (Lord, 1970), programmed testing (Cleary, Linn & Rock, 1968a) and, most recently, response-contingent measurement (Wood, 1972). Each of these terms attempts to convey the idea of adapting, individualizing or tailoring the testing strategy to a given individual based on responses he has made to test items already presented.

Research Approaches to Adaptive Testing. Several research strategies have been brought to bear on the question of whether ability tests should be adaptive or conventionally administered. Each type of study has its unique limitations and, because the kinds of generalizations drawn from the various kinds of studies are inherently limited by the approach taken, each study to be summarized below will be clearly identified by its basic strategy.

Empirical studies are, of course, a primary source of information on adaptive testing. These studies are characterized by 1) use of real people as testees; 2) use of a real item pool; and 3) administration of the ability test in a specified mode. Modes of test administration in empirical studies have included paper and pencil administration and the use of punch-board administration devices; administration by specially designed testing machines; and use of time-shared interactive computer systems to administer ability tests adaptively.

Conclusions drawn from empirical studies must be considered carefully, however, due to characteristics of the subjects being tested, the adequacy of the item pools being used, and the interactions of subjects and modes of administration. Some of these limitations of empirical studies will become more apparent following their discussion below.

Because of some of the difficulties encountered in empirical studies, and the limits of generalizability of these studies, a number of researchers have turned to simulation studies. This approach can be further divided into "real data" simulations and Monte Carlo simulations. "Real data" simulation studies use existing test data from the administration of conventional paper and pencil tests. These data include item responses of a number of individuals, total scores, and data on the difficulties and discriminations of the test items. To simulate adaptive testing on that group of subjects, the researcher adopts some adaptive testing strategy or strategies and re-scores each individual's answer sheet <u>as if</u> the test had been administered adaptively. The approach is, therefore, characterized by 1) real subjects, 2) responding to a pool of real items, but 3) under simulated adaptive testing strategies.

Conclusions drawn from "real data" simulation studies are, of course, limited by the nature of the item pool and the characteristics of the subjects. Although they are not limited by subject-mode interactions, they do lose valuable information on the actual effects of adaptive testing on the testee.

Monte Carlo simulation studies involve the generation of hypothetical item pools and hypothetical groups of subjects. In these studies, the investigator specifies exactly the characteristics of the item pools, in terms of item difficulties and item discriminations, and the ability levels of the "testees". Then, using item characteristic curve theory and computer-generated random numbers, matrices of testee "responses", total scores, and ability levels are generated for a pre-determined item pool, specified adaptive (and conventional) testing strategies, and a given number of subjects. While these kinds of studies obviously control for characteristics of the item pool and for the ability levels of the subjects, they provide no information about the actual effects of adaptive testing on testees. They do provide valuable information on the effects of item pool characteristics on results obtained by adaptive (as well as conventional) testing, but they are, of necessity, limited by the assumptions made in generating the test response records for simulated testees.

Closely related to the Monte Carlo simulation studies are the theoretical studies. These studies are based solely on item characteristic curve theory with items of specified characteristics, in terms of difficulties, discriminations, and guessing parameters. These studies differ from the Monte Carlo simulation studies in that they investigate a continuous range of hypothetical ability levels, rather than a pre-specified sub-set of abilities, for a theoretically "optimal" set of test items, and are not limited to a pre-specified number of simulated subjects. All results to date derived from theoretical studies are based on the solution of a series of mathematical equations due to Lord (1952; Lord & Novick, 1968) and others, which generate distributions of "test scores" from assumed item characteristic curves for "subjects" with varying amounts of assumed ability under a specified testing strategy. The results obtained from the solutions of these equations are, of course, completely dependent on the assumptions made in their derivation and on the assumed characteristics of the items. The studies are valuable, however, in that they permit the very rapid, but restricted, evaluation of a variety of testing strategies and parameters. As do the simulation studies, theoretical studies obviously do not permit the evaluation of the actual effects of adaptive testing.

The diversity of approaches to studying adaptive testing is, however, an indication of the extent of interest in the field. Comparison of results across a variety of types of studies permits a generality of conclusions that would not otherwise be possible. In addition, by following similar procedures with two different kinds of studies, sources of variance leading to different conclusions can be more readily identified. For example, administering a specified strategy of adaptive testing to live subjects in an empirical study and at the same time simulating responses to the same item pool under the same strategy might uncover subject-item pool interactive effects which could help clarify the conclusions derived from the empirical study.

Criteria for Evaluating Adaptive Testing. In addition to the diversity of approaches to studying adaptive testing, an understanding of the research in the area is further complicated by the different kinds of criteria on which adaptive testing procedures are evaluated. As might be expected, adaptive testing has been compared to conventional testing on practical criteria. These include such considerations as time involved in administration, cost of the various strategies of administration, and administrative complexity.

Some studies use as an evaluative criterion the correlation of scores on the adaptive test with scores on a conventional paper and pencil test. In these studies, conventional test scores are usually known in advance, and adaptive tests are either actually administered or simulated to obtain adaptive test scores, using items chosen from the conventional test. The objective in many of these kinds of studies seems to be to determine which strategy of adaptive testing best estimates the total score on a "parent" test. Studies using this approach tend to be either empirical or real data simulation studies.

A number of theoretical studies have used the correlation of test scores with underlying ability. A variation of this is found in the Monte Carlo simulation studies in which the criterion for evaluating adaptive testing strategies may be the correlation of generated or underlying ability with ability as estimated from the generated item response patterns for the hypothetical individuals. In these studies the researchers are interested in the "validity" of the testing strategies as the scores generated predict hypothetical "ability" for a group of hypothetical subjects.

A fourth criterion for evaluating testing strategies is by the use of "information functions." Rather than resulting in a single numerical index which reflects the relationships between two testing strategies, or the "validity" of a given strategy, the information function compares two or more strategies of testing in terms of the amount of information they provide at different levels on the underlying ability continuum.

The most prominent information function used in the literature on adaptive testing is due to Birnbaum (1968). Computation of Birnbaum's function results in a numerical value for each level of underlying ability, for a given testing strategy. The results are frequently displayed in a bivariate graph with underlying ability on the abscissa and information on the ordinate. Since the information values are interpretable only in a relative fashion, information graphs always compare two or more strategies of testing.

Birnbaum's information function can be interpreted in three ways. First, it reflects the relative number of items necessary for two tests to achieve the same level of precision of measurement. Thus, for a specified level of underlying ability, information function values of 20 and 10 respectively for testing strategies I and II indicate that strategy I provides 20/10 or 2.0 times the information as strategy II. Thus, Strategy II would require twice as many items as strategy I to achieve the same degree of precision of measurement.

One formula for computing Birnbaum's information function involves two terms: the numerator is the squared slope of the regression of observed test scores on underlying ability, while the denominator is the conditional variance of test scores at each value of underlying ability. The square root of the information function gives the reciprocal of the confidence interval for estimating underlying ability from observed score (Green, 1970); the information function therefore can reflect the "precision" of measurement at varying levels of underlying ability. Thus, a low value of information represents a large variance of observed test scores around true underlying ability (or a large standard error of measurement) while a large value of the information function represents a small variance of test scores around ability scores, or a small standard error of measurement.

Lord (1971a,d) presents a third interpretation of the information function. According to Lord, given two different levels of underlying ability, the information function represents the capability of observed test scores to discriminate the two levels of true underlying ability. This variation of the formula appears as a t-ratio type of statistic which has as its numerator the difference in means of observed test scores at the two specified levels of underlying ability and as its denominator the conditional variance of test scores for the two pooled levels of ability. Large values of the function indicate that test scores are very successful in differentiating the two levels of underlying ability, and small values indicate that the observed test scores do not clearly discriminate the two levels of underlying ability.

The three interpretations of the information function are interchangeable. Thus, values of information refer to 1) the relative number of items to achieve the same degree of reliability; 2) the relative standard errors of measurement; and 3) the relative capabilities of testing strategies to provide discrimination between levels of ability.

## RESEARCH ON ADAPTIVE TESTING

### Two-stage Procedures

Two-stage testing procedures are the simplest of the adaptive testing models. The two-stage strategy typically consists of a routing test followed by a series of "measurement" tests. The routing test is usually a set of items distributed across the ability continuum; its purpose is to make an initial estimate of each individual's ability level within a band of ability scores. Thus, the routing test might categorize individuals into 2, 4 or 10 initial levels of ability. Once a score has been determined for an individual on the routing test, and his ability classification estimated, an appropriate "measurement" test is selected for him, based on his ability classification, as his "second stage" test. The typical "measurement" test is a peaked test, consisting of a number of items all around the same level of difficulty. The level of difficulty of each of the second stage measurement tests, of course, varies. The testee therefore takes the routing test and only one of a series of measurement tests.

#### Empirical studies

The first reported study of two-stage testing procedures (and the only apparent empirical study) was by Angoff & Huddleston (1958). That study involved the comparison of two-stage testing procedures with conventional "broad range" ability tests on both verbal and mathematical abilities from the College Entrance Examination Board's Scholastic Aptitude Test. The two-stage procedure used a 40-item verbal routing test to route to two 36-item "high" and "low" measurement tests. For mathematical ability, a 30-item mathematical test routed to two 17-item measurement tests. All tests were timed. The study involved almost 6,000 students in 19 different colleges. The design was such that routing did not actually occur (i.e., the routing test was not scored prior to administration of the measurement test), but tests were administered in sufficient combinations to determine the effects of actual routing, had it occurred. Results of the study were evaluated in terms of reliability and validity considerations.

Results showed the narrow range (measurement) tests to be more reliable for the groups for which they were intended than wide-range tests, thus yielding scores with less error of measurement. Validities of the narrow range tests were found to be slightly higher than those of the conventional wide range tests, as measured against a criterion of grade point averages. Their data also show about 20% errors in classification due to routing.

Angoff & Huddleston (1958, p. 5) conclude that "there is a clear and relatively consistent superiority of each Narrow Difficulty-Range test to the Broad Difficulty-Range test in those regions of the ability continuum where both types of tests are appropriate," and that "a multi-level test offering for the College Board Program is technically superior, at least in terms of reliability and validity, to a single test offering." They do suggest, however, that the differences are not large enough, in view of the technical difficulties of an actual two-stage administration, to feasibly implement the routing test-measurement test procedure.

Simulation studies

The next series of studies of two-stage procedures appeared ten years later in inter-related papers by Cleary, Linn & Rock (1968a,b; Linn, Rock & Cleary, 1969). These studies were all "real data" simulation studies using the responses of 4,885 students to the 190 verbal items of the School and College Aptitude Tests and the Sequential Tests of Educational Progress. The total group was randomly split into a development and cross-validation group; several routing test procedures were developed in the development group and tested out on the cross-validation group.

Cleary et al. developed and evaluated four different two-stage procedures in their studies. Their "broad range" routing procedure consisted of a 20-item routing test with a rectangular distribution of difficulties as estimated on the total group. Based on an individual's score on this test, he was routed to one of four 20-item measurement tests consisting of items with high discriminations as estimated on a group with the same range of total scores, based on fourths of the total score distribution on the "parent" test. A second routing procedure used by these authors consisted of a double routing procedure, followed by one of the same four measurement tests. In the double routing procedure a 10-item routing test with items of average difficulty routed individuals to one of two second 10-item routing tests, each of which then routed to two 20-item measurement tests.

The third two-stage procedure used was referred to as a "group discrimination" procedure. In building this routing test, the score distribution of the parent test was divided into four intervals. Twenty items were then identified which had the largest between-group differences in item difficulties. The individual's total number correct on these 20 "group discrimination" items constituted his score on the routing test. He was then routed to a measurement test at the appropriate level of difficulty.

Cleary _et_ _al_. refer to their fourth routing test approach as a "sequential" routing test. In this method of routing, items would be administered to subjects one at a time. After each item response is determined, "likelihood ratios" are computed to determine an individual's likely membership in each of four ability groups. Given certain predetermined classification parameters, item administration in the routing test is terminated when the likelihood ratios permit a classification for each individual. The individual is then routed to the appropriate measurement test for his estimated ability level. In implementing this approach Cleary _et_ _al_. used both a three-group and four-group approach to the development of the sequential tests.

In these studies Cleary _et_ _al_. also studied two different ways of scoring the two-stage procedures. These methods included developing total scores both with and without use of the information obtained in the routing tests. For comparative purposes, two "best" conventional tests of 40 and 42 items were compared with the results of the two-stage procedures. Two papers (Cleary _et_ _al_., 1968a,b) report the results in terms of correlations with scores on the parent test, while one paper (Linn _et_ _al_., 1969) reports results as correlations with the "external criterion" of scores on the College Entrance Examination Board tests and scores on the Preliminary Scholastic Aptitude Tests taken one and a half years later.

Results of these studies showed that the sequential two-stage procedure correlated highest with total score. Next highest were the two conventional tests, followed by the group discrimination, broad range, and double routing two-stage procedures. The differences in correlations with total scores from highest to lowest accounted for only 6% of variance in total scores and were probably not statistically significant. Since the two-stage tests were typically composed of a much smaller number of items than the parent test, the authors suggest that the use of such procedures can achieve drastic reductions in the number of items administered to an individual with little or no loss in accuracy of total scores. Even the best short standard test was shown to require about 35% more items to achieve the same level of accuracy as the 3-group sequential two-stage procedure. Particular benefits in reduced time and increased accuracy would be expected for individuals who deviate from the average of the ability distribution.

The validity results were even more favorable for the two-stage adaptive procedures than were the correlations with scores on the parent test. In terms of the correlations with the "criterion" of other achievement and aptitude test scores, the group discrimination and 3-group sequential two-stage procedures achieved highest correlations. With the exception only of the double-branching two-stage model, the two-stage tests achieved higher validities than conventional tests of an equal number of items in every comparison; in most cases the validities of the 40-item two-stage tests were higher than those of the 50-item conventional tests. In five instances the 40-item adaptive tests correlated slightly higher with the external criterion than did the 190-item parent test, thus achieving equivalent validities with almost 80 percent fewer items. Linn _et_ _al_. (1969) conclude that "a test which was parallel to the 190-item total test would have to be 3.36 times as long as the best programmed test to have an equal median correlation with the outside criterion tests" (p. 145). It is important to note that these results were obtained by simulation of branched tests, as opposed to actual adaptive administration, which could be assumed to have additional advantages. Furthermore, the two-stage strategies were compared with an external criterion (other standardized tests) which could be expected to favor the standardized tests as predictors.

The results of these studies agree in general with those of Angoff & Huddleston (1958) although the differences are greater in the latter series of studies. Two-stage procedures appear to result in higher reliabilities, correlations with parent tests, and higher validities against external criteria. In both studies, about 20% of the testees were misclassified by the routing tests. This raises the question for future research on two-stage models of the effect of this mis-routing on the results. If the routing procedure had been recoverable, i.e., if the two-stage procedures were computer-administered so that routing errors could be detected and corrected before termination of testing, the adaptive strategies might have shown even greater advantages. A preliminary answer to this question could result from reanalysis of Angoff & Huddleston and Cleary et al.'s data, eliminating individuals for whom routing was in error.

Theoretical studies

Lord (1971e) has published the only theoretical study of two-stage testing procedures. His analyses are based completely on the mathematics of item characteristic curve theory under a specified set of assumptions. In his paper he reports on only the "best" results obtained from theoretical studies of over "200 different" two-stage strategies. His assumptions include 1) a fixed number of items administered to each "testee"; 2) dichotomous (right-wrong) scoring; 3) normal ogive item characteristic curves; 4) homogeneous items in a unidimensional test; 5) all items of equal discriminations, i.e., items differed only in difficulties; 6) both the routing tests and measurement tests were peaked, i.e., all items in each test were of the same difficulty; and 7) that all routing and measurement tests were linear (i.e., non-branched). The 200 different strategies studied varied in terms of total number of items (15 or 60), the distribution of items between routing tests and measurement tests (and, therefore, the number of levels of the measurement test), and whether or not random guessing was assumed (for a 5-choice item, within the 60-item studies only). Lord presents his results in terms of information functions, comparing the information obtained under the two-stage procedures with those obtained from a standard peaked test with all items of average difficulty.

Lord's results show that the best of his two-stage procedures provides almost as good measurement around the mean ability as the standard peaked test. As ability deviates from the mean, the two-stage procedures provide better measurement with the relative improvement increasing with increasing distances from the mean. Lord's best two-stage procedure was an eleven item routing test followed by six levels of measurement tests of 49 items each. Thus, each examinee would take exactly 60 items. Good results were also obtained for an 11-item routing test followed by four levels of measurement tests each with 49 items. Lord's results showed, however, that when guessing was assumed the measurement effectiveness of the two-stage procedures was seriously impaired, although it was still superior to the standard peaked test for the upper ranges of the ability distribution. Other aspects of his results give valuable suggestions for the future design of two-stage adaptive testing procedures.

## Summary

The empirical and simulation data on two-stage tests show higher reliability and validity for some of the two-stage procedures studied, with substantial savings in test administration time. Lord's theoretical results, while generally showing the capability of better measurement for two-stage procedures, are difficult to integrate with the other studies due to the different methodologies employed and the different criteria by which the results are evaluated. While the empirical and simulation studies are limited by the characteristics of the item pools used and by the characteristics of the subjects, they differ in many other respects from Lord's studies. For example, both Angoff & Huddleston (1958) and Cleary et al. (1968a,b) used routing tests which were not peaked, while Lord's (1971e) assumptions included routing tests of uniform difficulty. Lord's measurement tests differed only in terms of difficulty; Angoff & Huddleston's differed in both difficulties and discriminations; and Cleary et al's. were constructed on the basis of within-group discrimination values. Lord's results showed poor measurement for the two-stage procedures under random guessing; both other studies used real data on multiple-choice items on which guessing likely occurred, but without apparent detrimental effects on the results. Thus the results of these non-theoretical studies raise questions about Lord's assumption concerning random guessing.

In general, however, even in light of these differences in methodology and assumptions, the results of these studies seem to converge on the conclusion that two-stage adaptive testing procedures can give results as good as conventional procedures, and in many respects the accuracy and validity of measurement achieved by some of the two-stage procedures is superior. Two-stage procedures can also, in many cases, achieve this superiority with substantially fewer items than conventional ability tests.

## Multi-Stage Adaptive Testing

### Fixed Branching Models

Most of the research to date on adaptive testing has used the multi-stage model, rather than the two-stage approach. The most frequent applications of the fixed branching multi-stage model are based on a pyramidal or tree-structure arrangement of test items. These structures require an item pool which is ordered in terms of item difficulties. At the top of the pyramid consisting of the first stage of the multi-stage structure, is a single item which is typically of median difficulty for the group for which the test is intended. If the subject responds correctly to the first test item, he is typically administered an item of higher difficulty, moving down a right-hand branch of the pyramid to a second stage item; if the testee answers the first-stage item incorrectly, he is administered an item of lesser difficulty, moving down a left-hand branch of the pyramid. On the basis of the testee's response to the second stage item, he is "branched" to a left-hand or right-hand branch, respectively an item of lesser or greater difficulty. The branching process continues, typically, until a testee has responded to a test item at each of a number of stages. The pyramidal structure taken in cross-section at any stage beyond the first would show items in increasing order of difficulty going from left to right through the structure.

When a subject is to be administered one item per stage, there is one item available at stage 1, two items at stage 2, and n items at stage n. Each subject is then routed to one item at each stage based on his responses to the previous items. While these multi-stage fixed branching procedures require fairly large item pools for their implementation, the number of items actually administered to any subject (i.e., the number of stages) is fairly small. Typical multi-stage fixed branching studies use from 5 to 10 stages (respectively a 15-item and a 55-item pyramid), requiring each subject to respond to only 5 to 10 test items.

A number of variations of these multi-stage procedures have been reported in the literature. Some approaches have fixed entry points, typically an item of median difficulty. Others have proposed the use of variable entry points, tailoring the first item to be administered to an individual to be an item at his estimated level of ability, rather than an item of median difficulty for a group. The number of items to be administered at each stage also varies. In some studies as many as five items per stage have been used; others have used three. In these cases, differential branching occurs based on the number of items an individual has answered correctly at a given stage; in a three items per stage design the individual who answers all three items correctly is branched to an item of greater difficulty than the person who gets only 1 of 3 items correct. This kind of branching constitutes an implicit adaptive variation of "step sizes." The step size is the increment (or decrement) in difficulties from items at one stage to those at the next stage. Some studies use a fixed step size throughout; some use shrinking step sizes, with smaller changes in item difficulties at the later stages of testing to more adequately converge on the testee's ability level; and some studies use combinations of fixed and variable step sizes.

Another variation in the fixed branching studies appears in what has been called the "offset." The majority of studies use a "up-one, down-one" procedure, where a correct response on an item leads to an item one step higher in difficulty, and an incorrect response leads to an item one step lower in difficulty. Other studies, however, vary the offset so that a correct response to an item leads to an item one step higher in difficulty, while an incorrect response leads to an item 2 steps lower in difficulty; this is referred to as an "up-one, down-two" procedure, which may be used when guessing can be assumed to be operating.

Termination rules also vary among studies. The termination rule determines the number of items to be administered to a given subject. In most studies, the number of items to be administered is determined by the number of stages in the pyramid; however, in some cases it has been suggested that the number of items administered be controlled by determining when enough items have been administered to reach a desired degree of precision of measurement (Owen, 1969; Weiss, 1969; Wood, 1971), or when sufficient items have been administered to reach the decision for which testing is being implemented (e.g., Cronbach & Gleser, 1965; Ferguson, 1971; Green, 1970). Others (e.g., Lord, 1970) have suggested that testing cease when the range of item difficulties being administered to an individual centers around items of .50 difficulty for that person (i.e., he gets about 50% correct).

Scoring of fixed branching adaptive tests is accomplished in several ways, with different methods of scoring yielding different results. In some studies the score for an individual is the rank of the difficulty of the final item reached; thus, in a 6-stage pyramid, only six score values are possible. Others use the correct/incorrect information of the final item administered to obtain double the number of score ranks. In this approach a 6-stage model would yield twelve score values, since a correct or incorrect answer leads to two possible ranks for each of the six items. Some studies use the diffculty level of the final item reached, or extending the logic of the previous approach, the difficulty level of the "n + 1$^{th}$" item, to utilize the response information of the last item administered. Still others use the average difficulties of all items administered to a given testee, or a weighted average of item difficulties, giving more weight to the items administered to an individual later in the sequence.

It is clear that there are a very large number of combinations of approaches to fixed branching adaptive tests. Yet with all the variability used in entry points, step sizes, termination rules, and scoring schemes, as well as the differences in approaches taken by the empirical, simulation, and theoretical studies, the research to date does appear to converge on a common conclusion.

Empirical studies. Multistage branched testing was first reported in 1956 by Krathwohl & Huyser, using a modification of paper and pencil answer sheets to route students through one of two fixed branching adaptive tests. This study used an eight-stage, one item per stage model, and a four-stage, two items per stage approach. Based on a group of 100 college students, Krathwohl & Huyser obtained a correlation of .78 between their sequential test and the 60-item parent test, showing a large savings in testing time with only a moderate loss in the information obtained from the longer test.

Krathwohl & Huyser's work in paper and pencil sequential testing was extended by a group of Army researchers led by Bayroff (Bayroff, Thomas & Anderson, 1960; Seeley, Morton & Anderson, 1962). Bayroff's group developed four different 6-stage branched tests, one for each part of the Armed Forces Qualifications Test (AFQT). Like Krathwohl & Huyser, they used an up-one down-one approach, with decreasing step size and one item per stage. Entry point was constant at median item difficulty ($p = .70$), and score was the ranked difficulty of the $n+1^{th}$ item. One innovation introduced in Bayroff's studies was the use of differential branching on the first item for incorrect answers, based on the difficulty of the chosen distractor.

Bayroff administered his sequential tests by paper and pencil with the chosen answer giving the examinee the number of the next item to be taken; he included a number of unused "buffer" items to hide the routing sequence from the testee. Results of administering two of the branched tests to about 500 men were evaluated by a comparison of score distributions and correlations with total scores on the parent tests.

Results showed a correlation of .63 for the 6-item sequential test with the parent test. Conventional tests of 25 items correlated higher with the parent test than did the sequential tests. Further analysis showed that apparently the sequential tests were too easy; scores were badly skewed with definite bunching at the high score end of the distribution. This finding alone could account for the lower correlation of the sequential tests. The sequential tests also took considerably longer to construct, longer to administer than conventional tests of equivalent length, and resulted in more unusable sets of data than conventional tests, due to the testees' failure to follow the routing instructions. While scoring of the branched test was easier, since it involved simply determining whether one of a number of possible terminal items was correct or not, the verification of the routing process was considerably more time-consuming than required for scoring of conventional tests.

Similar negative results were found in a paper and pencil study of a branched test reported by Wood (1969). Wood developed branched tests of 4, 5 and 6 stages and administered them to 91 students. He used a fixed step size procedure, entry at median difficulty with an even offset, and scored them using total number of correct answers (varying from 0 to 4 through 0 to 6). His criterion was the correlation of test scores with course grades.

His results showed correlations of about .35 for the 4 to 6 item branched tests with course grade. When the three sub-scores from the three multi-stage tests were combined into a total score, that score correlated .51 with the course grade. The results also showed that scores on the conventional test and the score derived from the "best" 15 items in the conventional test were both better predictors of grades than were the scores on the branched tests or the score on all three branched tests in combination.

Wood's study has a number of deficiencies which limit the generality of his conclusions. First, it appears that the branched tests were selected to measure separate components of mathematical ability, while the conventional test included all three components in combination. Thus, a fair comparison of the two approaches as they predict a heterogeneous criterion would have required a heterogeneous branched test. Secondly, Wood did not report the distributions of scores on the branched tests. With the limited ranges of scores possible in tests of from 4 to 6 items, it is likely that the restricted range of scores and their possible skewness if the branched tests were poorly constructed could account for the low correlations with grades. Thirdly, the paper and pencil approach to administration of the branched tests could have resulted in additional error variance; use of a complex paper and pencil branching test can confound test scores by an extra component resulting from the testee's ability (or willingness) to implement the branching procedure, as it interacts with the ability being measured.

Because of the administrative problems involved in using multi-stage branched tests in paper and pencil format or variations of that format (e.g., specially designed punch boards), researchers have turned to mechanistic and automated devices to administer adaptive tests. Bayroff (1964) reports on the design of a "programmed" testing machine which can administer linear (conventional) tests, two-stage, multi-stage, and combinations of these ability testing strategies. The machine was designed to conserve testing time by terminating testing if a testee's performance fell below or above pre-specified points. In addition, the machine provided for 1) the capacity to permit the subject to choose a tentative selection of answers prior to deciding on one alternative multiple-choice response (a form of differential weighting of response distractors; 2) recording of response latency data; and 3) administration of immediate feedback to the subject on the correctness of his responses.

The testing machine Bayroff designed was apparently never put into production. In its place, the growth of time-shared interactive computer systems permitted Bayroff and others to continue research into adaptive testing, with different results from those derived from paper and pencil adaptive testing.

Bayroff & Seeley (1967) administered two eight-stage branched tests on a teletype connected to a time-shared computer (9 stages were used for the most able subjects). Their branched test included difficulty levels varying from .95 to .20, entry point at an item of .60 difficulty, and a fixed step size of .05. Test items were six distractor multiple choice measuring verbal and numerical abilities. Test score was the relative difficulty of the n+1th item, giving a score range of 17 points. The two branched tests were administered to 102 subjects who also completed a 50-item verbal test and a 40-item numerical test, both conventional tests composed of items from the same pool used to construct the branched test.

Analysis of the data yielded correlations (corrected for restriction in range) of .83 and .79, respectively for the verbal and numerical tests, with scores on the conventional tests. These correlations approached the test-retest reliabilities of the conventional tests (r = .91 and .85 respectively). Conventional tests of the same length were estimated to have correlations of .75 and .67, respectively, with the parent test (Bayroff, 1969).

Computer administration of the branched tests reduced the correlations between verbal and numerical tests from .65, which resulted from paper and pencil administration, to .57. Scores on the conventional test and the verbal branched test were approximately normal, while those on the numerical branched test were piled up at the high end of the distribution. Individuals with maximum scores on the latter test were distributed over two standard deviations on the similar conventional test (Bayroff, 1969). One possible explanation for this finding is that the adaptive administration, because it tailors items to the individual's ability level, permits more individuals to obtain "true" high scores by eliminating sources of error variance in conventional test administration which artifactually depress test scores for certain testees.

A major conclusion derivable from Bayroff & Seeley's study is that conventional linear tests would have to be about twice as long as the branched tests to achieve the same correlation with the criterion paper and pencil test. Thus, adaptive computer administration of ability tests appears to require about 50% less items (and, therefore, shorter testing times) to achieve the same amount of information, based on the criterion used in this study.

Hansen (1969) also administered an adaptive test by teletype. He used achievement test items in five 3- to 5-stage pyramidal subtests, so that the total test consisted of 17 items per individual. Hansen's pyramid used an entry at p=.50, step size of .10, and a variety of scoring methods based on final level of difficulty reached. The 56 students who completed the adaptive tests had also taken a conventional achievement test on the same material one week earlier. Scores on another achievement test and course grades were used as criterion variables. In addition, special reliability indices were computed for the computerized test and compared to the reliabilities on the 20-item conventional test.

Analysis showed that at least one approach to scoring the computerized test yielded 1) a more rectangular score distribution than did the standard test, which yielded a skewed distribution; 2) higher subtest reliability and higher total test reliability than the 20-item conventional test; 3) shorter testing time; 4) higher correlation with final grade; and 5) higher correlation with the achievement test criterion. These findings were replicated in a second study which also showed college freshmen to have positive attitudes toward computerized testing.

A third study of computerized branched testing was reported by Bryson (1971). This study compared two special branched procedures with results from two short conventional tests. Her criterion was correlation with total scores on a 100-item parent test. Paper and pencil tests were 5-item tests in which items were selected using special item analysis techniques. Branched tests were administered on a cathode ray computer terminal with response by light pen. The branched tests each consisted of five stages (with a sixth item for those who correctly responded to the most difficult item); items were arranged in variable step-size order. The branched tests were constructed using a variation of Rasch's (1966a,b) item analysis model and a specially designed item selection approach which sequentially selects items for a pyramidal structure based on the most valid item for all individuals who reach a given node in the pyramidal routing procedure. Computerized tests were administered to two groups of 263 testees, while the conventional tests were administered to 250 individuals.

Bryson's empirical results are not generally in favor of the computerized administration. Correlations of computerized test scores with scores on the parent tests were virtually identical with those of the 5-item conventional tests. However, a careful analysis of the branching paradigm for one of her adaptive strategies shows that one item selection technique did not place items in a meaningful order of difficulties; another of her pyramids had a very restricted range of difficulties. Furthermore, distributions of scores are not given for any of her results, making it impossible to determine if a truncated or skewed score distribution, such as found by Bayroff & Seeley (1967), could account for her findings. Another limitation of Bryson's results derives from her method of scoring the branched tests. "Scores" on Bryson's tests were obtained by identifying each possible pathway through the branched test and determining, in a developmental sample of 10,000 recruits, the mean total score on the parent test for those with the same pattern of response. No indication is given of the number of subjects on whom each of these means were based, thus scores are of unknown reliability. This procedure also assumes the inherent similarity of adaptive and coventional test administration, an assumption which should be called into serious question and which might, in part, account for her results.

The empirical studies available to date on fixed branching models show mixed conclusions. In general, when well-designed adaptive tests were studied it appears that branched adaptive tests show promise of effecting considerable savings in test administration time, through the use of fewer items, than conventional tests. Two computerized test administration studies agree in showing different distributions of scores under computerized than paper and pencil administration, while Hansen (1969) reports higher validities for computerized test administration than for conventional administration.

Simulation studies. Two "real data" simulation studies report results for fixed branching multi-stage adaptive testing. Bryson's (1971) study compared her empirical results with results on the same testing strategies based on available item response data from two groups of 100 recruits. These analyses showed one of the branching strategies to have consistently higher correlations with total test score than either the two conventional strategies or the other branched strategy. The second branched strategy had lower or equal correlations than one conventional strategy and higher correlations than the other. These results contrast quite clearly with the empirical results, which showed equal correlations for all methods. The differences suggest caution in drawing conclusions from simulation studies and generalizing them to empirical studies; apparently the actual process of administering an adaptive test might have effects which do not occur in simulation of adaptive administration from data already administered in conventional testing formats.

Linn et al. (1969) in their study of two-stage models also used available conventional test responses to simulate administration of two multi-stage strategies. One of their tests was a 10-level pyramidal model with entry at p=.65, step size of about .02, and an equal offset (up one/down one). Test scores were based on the addition or subtraction of step size to a constant following a correct or incorrect response. Their second test was a 5-stage branched test with five items per stage. Branching occurred on the basis of an individual's scores at each level; scores of 0, 1 or 2 branched to an easier group of items while scores of 3, 4 or 5 branched to a more difficult group of items. Item difficulties varied slightly within each group of 5 items and step sizes between levels varied somewhat. As in the 10-stage test, total scores were derived by adding or subtracting .05 (the average step size) to a constant after each upward or downward branching, respectively.

Results showed that the 10-stage branched test had the lowest correlation with total score of the two multi-stage strategies, all the two-stage strategies, and the short conventional tests. It should be noted, however, that all items in the experimental tests were selected from the items in the parent test. Further, the results reported by Linn et al. show that the correlations with total score were roughly proportional to the number of items in the tests. Hence, the fact that the 10-stage branched test correlated lowest with total score could be partly explained by the fact that it had fewer items in common with the parent test than any of the other tests, except the 10-item linear test. Results for the 5-stage branched test (in which 25 items were "administered" to each testee) showed correlations with total score about equal to those of a 30 to 40 item conventional test. Thus, 25 items were used in a branched strategy to extract about as much information as a 35-item conventional test.

Of the adaptive strategies studied by Linn et al., the 10-stage branched procedure correlated lowest with the external criteria used. However, with number of items administered held constant, the multi-stage adaptive procedures correlated higher with the criteria than conventional tests of equal length, as did the two-stage procedures. The five-stage branched test (25 items) had correlations with the criterion tests higher than those of the conventional 50-item tests. These data suggest that multi-stage branched tests, as well as the two-stage models studied by these investigators, can result in considerable time savings in test administration with gains in validity, as compared to conventional tests.

An early monte carlo simulation study by Paterson (1962), deriving from Krathwohl & Huyser's (1956) pioneering work, provides additional information on the characteristics of fixed branching multi-stage models. Paterson studied a six-stage pyramidal test in comparison with a 6-item conventional test. His entry point was an item of 50% difficulty, and his branching rule chose a more difficult item following a correct response and an easier one following an incorrect response.

Paterson's step size rule is perhaps unique in research to date on adaptive testing. In constructing his item pyramid, Paterson ordered his items by difficulty and, within difficulty levels, by discriminations. Thus, the first items administered were the most discriminating and the last least discriminating at a given difficulty level. Step size varied as a function of item discrimination; a larger step increment followed a correct response to a highly discriminating item and a smaller increment for a correct response to a less discriminating item. Since items were ordered in terms of discriminations, the procedure approximates a "shrinking step size" procedure, with larger steps taken for early items and shorter steps for later items. Paterson's score on the branched test was the difficulty level of the final item administered.

Paterson generated a hypothetical population of 1500 "testees", 100 at each of 15 ability levels. Item discriminations varied, using biserial correlations of .45 to .79. Paterson assumed that guessing did not occur. He compared the sequential and conventional tests under conditions of normal, rectangular and U-shaped ability distributions, as well as similar score distributions.

Results of the study showed that the branched test better reflected atypical (e.g., U-shaped) ability distributions in test scores. The branched test also gave more precise test scores, particularly at the extremes of the ability distribution, since it more accurately classified individuals who were at the extremes of the ability distribution. Paterson also noted that both tests were about equal in the accuracy in which they, overall, predicted ability from test scores. Thus, Paterson's results suggest that the criterion used to compare the adequacy of the methods may have a direct effect on the conclusions drawn. As a subsidiary, but important, finding Paterson observed that the sequential tests were not sensitive to errors in estimating the item parameters.

While the three simulation studies vary widely in approach, subjects, testing strategies, and evaluative criteria, the results are generally in favor of adaptive testing. Bryson's (1971) study shows one adaptive approach to be superior to conventional procedures in terms of correlation with a parent test. Linn et al.'s (1969) data shows the branched tests to have considerably higher validity, with number of items held constant, than conventional tests. And Paterson's study, although it does not yield higher correlations with underlying ability for the branched test, does show the branched test to be more sensitive to distribution of underlying ability and to yield scores that are more precise than those of the conventional test.

Theoretical studies. A number of investigators have studied fixed-branching multi-stage models using mathematical derivations from item characteristic curve theory. In 1964, Waters (under Bayroff's direction), reported a theoretical study comparing a 5-item conventional test and a 5-stage pyramidal adaptive test. The conventional tests were developed in four different forms to reflect different spreads of item difficulties. The sequential test used an up one/down one branching rule with increments of .10 in difficulty levels and final score as the difficulty level of the n+1th item. Both tests fixed item discriminations at .80, no guessing was assumed for some of the analyses, and fifteen levels of underlying ability were studied. The criterion in this study was the correlation of test score and underlying ability.

Results showed the correlation between test score and ability to be higher for the branched test than for the conventional test, even when random guessing was assumed. Additional analyses showed that the branched test, using final difficulty score, had a flatter score distribution (and, therefore, scores of more nearly equal precision) than did the conventional test.

Waters & Bayroff (1971; Waters, 1970) report a similar study extending these findings. In this study, they compared branched and conventional tests of 5, 10 and 15 items. While the branching models were basically the same as in the earlier study, they also studied a 2-items per stage multi-stage adaptive test. In this study, point-biserial correlations of items with the underlying continuum were systematically varied from .30 to .90 on the 5 and 10 item tests and from .40 to .80 on the 15 item tests. They also used 29 ability levels, ranging from +3.5 to -3.5, to study the results at all practical ranges of the ability distribution.

Using, again, correlation of test scores and underlying ability as the criterion, their results generally showed higher correlations for the branched tests than for the conventional tests, particularly at higher point-biserial correlations. Thus, when items are more discriminating, test scores on branched tests more accurately reflect "true" position on the underlying ability continuum. Comparison of the one-item per stage and two-item per stage branched tests showed no improvement for the latter strategy.

In a series of interrelated papers, Lord (1970; 1971a,e) has presented a considerable amount of theoretical information on the characteristics of fixed-branching adaptive testing models. A brief but incomplete overview of his method and results is given in Lord (1971c); the theoretical basis is in Lord (1972).

All of Lord's analyses are evaluated in terms of Birnbaum's (1968) information function. It will be recalled that this function reflects, at each level of underlying ability, a value based partially on the precision of measurement, related to the standard error of measurement, at that ability level. Higher precision implies a lower standard error and lower precision a higher variability of observed scores around true scores.

All of Lord's theoretical analyses, with only minor exceptions, are based on a common set of assumptions. These include 1) normal ogive item characteristic curves; 2) all items of fixed and equal discriminating power (biserial correlations of about .45); 3) items that vary only in difficulties; 4) a fixed number of items to be administered under the branching strategy; 5) either no guessing or completely random guessing; and 6) a comparison peaked conventional test with all items having equal difficulties (the mean of the population being tested) and equal discriminations.

Lord (1970, 1971a) and an associate (Stocking, 1969) studied a variety of tailored testing strategies using 10-stage, 15-stage, and 60-stage procedures, although not all strategies were studied for each size branched test. Strategies studied include equal step size procedures using branching rules of up one/down one, up one/down two or three, with different constant step sizes, based on a Markov chain random walk model. Lord also studied Robbins-Munro shrinking step size procedures, based on a mathematical model adapted from work in bioassay. In addition, he studied other shrinking step size procedures designed to approximate the Robbins-Munro procedures, but without making the same formal assumptions. In some of his studies, Lord compared several scoring procedures for tailored tests, including average difficulty score and final difficulty score. In all his studies, the entry point, or first item "administered" under tailored testing procedures, was always an item of average difficulty for the group being tested.

Because of the variety of strategies studied, Lord's results are difficult to summarize. However, one finding is fairly clear. Under the assumptions from which the results were derived, the conventional test always provides more accurate measurement than any adaptive strategy at the mean of the ability distribution. Thus, the information curve for the conventional test approximates a normal curve, highest at the mean and dropping off sharply as ability deviates from the mean in either direction. Information curves for the "good" tailored tests, however, do not have the bell-shaped characteristic. Rather, information curves for adaptive strategies approximate a horizontal line, crossing the information curve for the standard test between .5 and 1.0 standard deviations on the ability distribution and remaining relatively flat out to at least ±3.0 standard deviations. Thus, while the precision of the conventional test is highest at the mean of the ability distribution, the good adaptive testing procedures give almost constant precision throughout the ability range as a result of administering items which are as closely matched to an individual's ability as is possible.

Following his analysis of the Robbins-Munro procedures, Lord (1971a, p. 14) concluded that "tailored procedures provide good measurement for a much wider range of examinee ability than does the standard test." Stocking (1969, p. 5) reached a similar conclusion for 15-item tests under a Robbins-Munro procedure. And, in his study of fixed step size procedures, which also included a comparison with a typical unpeaked "published" test, Lord (1970, p. 179) concluded that tailored testing is better than the "published" test for examinees at all levels of ability. In general, Lord's data show good tailored tests to provide better measurement for about two-thirds of the typical ability range, or about 30% of a normally distributed population, with larger percentages possible depending on the distribution of ability in the population.

The specifics of Lord's findings vary, of course, depending on the tailored strategies. In general, his analyses show that tailored tests lose some of their efficiency when random guessing is assumed. Under these circumstances, information functions become asymmetric and, under certain tailored strategies, the nearly constant precision of the tailored test is lost. Comparison of the utility of final difficulty scores with average difficulty scores shows average difficulty scores to be superior. Among the fixed step procedures, Lord found the up one/down one procedures to be more efficient than those with a variable offset, except when guessing was assumed. Step size itself had substantial effects on amount of information obtained under tailored testing.

Lord's results show the shrinking step size Robbins-Munro procedures to be superior to the fixed step size procedures. When compared with two-stage testing procedures (Lord, 1971e), the fixed step size procedures are about as good as the two-stage procedures (with number of items equal), but the multi-stage models provide greater precision at the extremes. Neither, however, is as good as the Robbins-Munro procedures, but both are better than non-Robbins-Munro reducing step size procedures.

In addition to using theoretical derivations to study some conventional multi-stage fixed branching adaptive models, Lord (1971b) developed a new multi-stage branched technique which he calls a "flexilevel" test. A typical multi-stage pyramidal branched test has at each stage a number of items, one of which will be administered to a testee based on his response pattern on previous items. This results in there being available for administration two items at stage 2, three items at stage 3, and so on, so that a 60-stage fixed branching test will require that there be available 1,830 items, of which only 60 will be taken by any one testee. Lord's flexilevel test, however, does not make such heavy demands on an item pool. A 60-stage flexilevel test, in which any individual will complete 60 items, requires an item pool of only 119 items. Lord accomplishes this by administering, following a correct response, the next more difficult item previously unanswered and, following an incorrect response, the next easier item previously unanswered. Each person continues answering until he has answered exactly half the items in the flexilevel test. As proposed by Lord, the testing procedure is paper and pencil and requires that the answer sheet inform the testee of the "correctness" of his response for routing to the next test item. The procedure is designed so that all individuals who arrive at a given terminal item have taken exactly the same items, in contrast to the typical pyramidal branched test in which a variety of pathways are possible to a given terminal item.

Lord (1971d) presented some theoretically derived data concerning his flexilevel test. Consistent with his previous analyses he assumed a 60-item flexilevel and a 60-item conventional test, both with constant item discriminations. He also compared the information functions for both tests with the information derivable from a test designed to discriminate at two points on the ability continuum. His results showed that the flexilevel test provides more information throughout the ability range than does the test designed to discriminate at two points on the continuum. The conventional peaked test, of course, provides more accurate measurement around the mean ability level, with the flexilevel test becoming more accurate at more extreme ability levels. Consistent with his previous theoretical results, the flexilevel procedure provides greater accuracy of measurement for at least 30% of the population, those who deviate beyond ±1 standard deviations on the ability distribution. However, the information curves for the flexilevel test are not as flat as those for the Robbins-Munro procedures, or for the best pyramidal procedures, thus showing less precision of measurement for the flexilevel test at the extremes of the ability distribution.

Another approach to reducing the demands of the multi-stage strategies on item pools was taken by Mussio (1972). He modified Lord's model to assume a Markov chain with a retaining barrier and a reflecting barrier. This modification involves truncating the upper and lower tails of the item pyramid, eliminating all items above and below specified difficulty levels. Thus, rather than having, say, 15 items available at the 15th stage of the pyramid, Mussio's approach might have as few as 11 items available. In the retaining barrier approach, testees reaching the highest or lowest difficulty level continue to receive items at that level; the reflecting barrier method would alternate between items at that level and available items at the next lower level (in the case of difficult items) or higher level (in the case of easy items). For a 60-stage truncated pyramid allowing a maximum of 11 difficulty levels, Mussio's approach requires only 262 items for the reflecting barrier and 390 items for a retaining barrier (compared to Lord's requirement of 1,680 items for a complete pyramid). Thus, this approach results in reducing item pool requirements by over 75%.

Mussio's theoretical analyses, presented in the form of information curves, show results similar to those obtained by Lord. In comparison to the peaked conventional test, adaptive tests provide less information at the mean of the distribution, but considerably more information for individuals whose abilities deviate from the mean. His comparison of the retaining barrier and the reflecting barrier showed the retaining barrier to maintain more nearly equal precision throughout the range of abilities than the reflecting barrier; however, both approaches showed some reduction in precision at very extreme ability levels, although both were still considerably more precise than the peaked conventional test.

Summary. Studies of both multi-stage and two-stage fixed-branching adaptive testing procedures have used empirical, simulation, and theoretical procedures to examine the characteristics of the pyramidal branching models and their derivatives. These studies have used a variety of item pools, both real and simulated, a variety of subjects, and have varied such characteristics of the adaptive testing procedure as step size, offset, and constancy of step size. In addition, the criteria on which the outcomes are evaluated have varied from study to study.

In general, the results show a definite advantage for adaptive tests in terms of number of items to be administered to any individual. Multi-stage branched tests that are well-designed (e.g., Bayroff & Seeley, 1967; Krathwohl & Huyser, 1956; Linn et al., 1969) give higher correlations with parent tests than do conventional tests of the same length, resulting in shorter testing times (Hansen, 1969). Similar results were found in Bryson's (1971) simulation study, at least for one branched procedure. Adaptive tests also give higher correlations with external criteria (Hansen, 1969; Linn et al., 1969), requiring conventional tests to consist of up to twice the number of items as multi-stage adaptive tests to achieve the same external validity. Multi-stage branched tests also give different distributions of test scores than do conventional tests (e.g., Bayroff, 1969; Waters, 1964), with these distributions better approximating an equi-discriminating rectangular distribution (Hansen, 1969) and better reproducing atypical distributions of underlying ability (Paterson, 1962), than do conventional tests. Finally, multi-stage branched tests give scores which, with highly discriminating items, have higher correlations with underlying ability for a fixed number of items than do conventional tests (Waters & Bayroff, 1971) and yield scores with more nearly constant precision of measurement and considerably greater precision of measurement for individuals at ability levels divergent from the estimated average ability of a group (Lord, 1970, 1971a, d,e; Mussio, 1972; Paterson, 1962).

## Variable Branching Models

As described above, the fixed branching multi-stage adaptive testing models use a structured item pool in which items are placed for administration in pre-determined order, based on their difficulty and discrimination parameters. Furthermore, in the fixed branching models there is a pre-determined step size which is constant across all individuals; even the shrinking step size procedures do not adapt to individual differences. The fixed branching models always depend on a pre-determined branching rule which determines whether the next item will be an item of higher, lower, or equal difficulty. In fixed branching procedures the number of items administered to an individual is also usually fixed in advance.

In contrast to the fixed branching models, the variable branching models require simply an item pool with known characteristics rather than a structured item pool. For the variable branching models items need only to be identified by appropriate indices of difficulty and discrimination; they are not organized into a hierarchical structure, nor need they be stratified according to difficulties or discriminations. The variable branching models do, in general, need to assume a specific (e.g., normal) distribution of ability in the testees.

In general, the variable branching models require the ready availability of computers for their implementation. The general procedure consists of choosing each item in succession for each individual, based on his responses to all previous items, in order to maximize or minimize some measurement-dictated criterion for that individual. Testing usually continues until some pre-specified value of the criterion is reached. Each item is selected by searching through the entire item pool of unadministered items to locate the next "best" item for that individual. While research with variable branching models has been sparse, both Bayesian and non-Bayesian approaches have been reported.

Bayesian strategies. Novick (1969) develops a Bayesian adaptive testing model based on classical true and error score theory. His model uses a regression-based approach based on the availability of a large and diversified homogeneous item pool. Novick's model uses both information available on the individual and information available on the population of which he is a member. In the early stages of testing, where only a few items provide a small amount of information on the individual, the weighted Bayesian regression model uses the mean of the population to provide most information. In the later stages, when a larger number of test items provide more specific information about the individual, his item responses are weighted more and the population data is weighted less. Items to be administered to an individual at each stage are based on a weighting of the individual's test responses and the population mean test score.

Novick's procedure for item selection is designed to find an item that has a difficulty level such that all people with a given ability level have a probability of .50 of obtaining a correct answer. This is the item, as suggested by Hick (1951), which provides the most information about the testee's ability. Using the prior ability estimates based on testee responses and population means, the Bayesian estimation procedure continually updates the ability estimates and provides information on which an individualized step size is chosen for the next item. Novick suggests that the Bayesian procedures will be especially valuable for short tests (15 to 20 items), a finding which is later remarkably well supported by Wood's (1972) results using a different Bayesian procedure.

Owen (1969, 1970) presents a Bayesian adaptive testing procedure different in a number of respects than Novick's. His model does not use information on group membership to arrive at ability estimates but bases all calculations on the item responses of one individual in an ability testing situation. The model assumes dichotomous (correct/incorrect) responses, local independence of item responses for any individual (i.e., all responses determined solely by underlying ability), normal ogive item characteristic curves, and a normal distribution of underlying ability. Owen develops the model under both guessing and non-guessing assumptions. Implementation of the method requires a prior estimate of the individual's ability and, therefore, permits a variable starting point for testing.

Owen's model is based on estimating a "loss function" at each stage of testing. Once the loss function, which is related to the "seriousness" of errors of estimating ability, is specified the choice of a scoring function and sequential decision criteria can be determined. Owen uses a quadratic loss function, which has the effect of reducing the variance of the ability estimate at each stage in the item administration procedure. The procedure chooses for administration to a given individual that unanswered item among all remaining items which minimizes the expected posterior loss. The item is then administered, and the new ability estimate and its variance are computed. This new posterior ability estimate then becomes a Bayesian prior ability estimate, and a new test item is chosen to reduce the next posterior loss estimate. That item is administered, the posterior estimates calculated, and the new prior is formed. The process continues until the variance of the posterior ability estimate, or the precision of that ability estimate, reaches a prespecified value. Owen develops approximation procedures for choosing items since the ideal test item might not exist in an actual item pool.

Owen's model has considerable inituitive appeal. For example, under his assumptions the posterior variance of the ability estimate is always smaller than the prior variance, even if the item is answered incorrectly. In other words, any item provides some information, permitting the procedure to "converge" more accurately on actual ability level. In his development using the guessing parameter, the relationship between ability estimate and difficulty is curvilinear, with the estimate of ability increasing with difficulty up to a point, but beyond that as an item becomes too difficult for an individual, decreasing the estimate of ability. Or, in other words, when an item is near an individual's ability level, he is less likely to guess, but as the item becomes more divergent from his true ability, random guessing is more likely to occur and to artificially inflate his observed test score. These assumptions about guessing stand in contrast to the "across the board" random guessing assumptions employed by Lord.

Wood (1971) programmed Owen's Bayesian model for actual test administration. He administered, on a time-shared computer, a pool of vocabulary items to 28 school children in grades 4 to 6. Subjects were required to continue answering each test item until they obtained a correct answer and, of course, were told when the answer was correct. Wood administered an average of 50 items per subject and studied the mean and variance of posterior ability estimates as a function of number of items administered. He also did some additional simulation studies using 1) the characteristics of the real item pool but simulating subject responses based on item characteristic curve theory, and 2) simulating both subject responses and item characteristics. Wood compared his results to a simulated two-stage approach, with 10 items at the first stage and 50 items at the second, and with a 60-item simulated conventional test.

Wood's results with live data showed that for a number of subjects the Bayesian ability estimates converged at around 20 items, as predicted earlier by Novick (1969) using a different Bayesian model. Thus, about 85% of the error reduction had occurred by item 20, on the average. In some cases convergence occurred much earlier, in some cases later, with the results partly based on the adequacy of the prior ability estimate for a given individual. In his "real item" simulation studies, Wood found the two-stage procedure best, followed by his Bayesian procedure and the conventional test, although he found some person-test interactions suggesting that some testing procedures might be more appropriate for some individuals than others. Using simulated item responses and a simulated item pool, Wood replicated the finding that the Bayesian procedure required only 20 items to effect 85% reduction in error. He also found that even with one-third fewer items the effectiveness of the Bayesian procedure matched that of the two-stage and conventional testing procedures. Thus, important savings in number of items administered to testees are evident from the Bayesian test administration procedure.

<u>Non-Bayesian strategies</u>. In a study designed to test the robustness of logistic test models, Urry (1970) developed an adaptive testing strategy which does not use a fixed branching approach. It is similar to the Bayesian methods in that items to be administered at later stages in the testing process are chosen in order to minimize the standard error of the ability estimate from the testee's responses to a given sequence of items. His method, however, is not based on Bayes theorem. Rather, it uses maximum likelihood estimates at each stage of the testing process to estimate

ability and its associated standard error, based on test items already administered. Urry's method bears some similarities to the fixed branching approaches in that the first item administered is an item of median difficulty. The response to that item determines a fixed branching for the second item, which is the most difficult item available following a correct response, and the least difficult item available following an incorrect response. Once an individual's response pattern deviates from all correct or incorrect, Urry begins his estimation procedure and moves to the variable branching model.

Urry's monte carlo simulàtion study compared conventional and adaptive tests under two models; Rasch's (1966a,b) 1-parameter model, in which guessing is not assumed and items differ only in terms of difficulties (the same model studied by Lord), and a two-parameter variation of the same model in which guessing is assumed. In contrast to Lord's studies, Urry systematically varied 1) item-ability biserial correlations from .45 to .85 in steps of .10; 2) item difficulties, using a constant value, normally distributed difficulties, and rectangular difficulty distributions; 3) guessing probabilities of .00, .25 and .50; 4) number of test items, from 10 to 50 in steps of 10; and 5) in a subset of studies item discriminations were unequal to study the effect of departures from the assumptions of the model. In all, Urry generated 36 different kinds of item structures. He calibrated his item banks on 500 hypothetical subjects and carried out all validity computations on an independent cross-validation sample of 100 "subjects" of known ability. His criterion for comparing methods and item banks was the validity correlations of known ability estimate with ability estimate derived from the model applied to the pattern of item responses of the cross-validation group.

Urry's results offer some suggestions for the design of adaptive tests, at least of the type he used. He found adaptive tests to increase in validity with increasing item discrimination, particularly for rectangular difficulty distributions; when item discriminations were high, a 10-item rectangularly distributed tailored test is as good as a 30-item peaked tailored test. When item discriminations varied, a rectangular distribution was also found best. He also found that his tailored testing procedure was adversely affected by guessing probabilities of .50, suggesting that his type of adaptive testing is not appropriate for true-false tests.

In comparing adaptive and conventional tests Urry's data show that tailored testing gives higher validities than a peaked conventional test when 1) the model is approprate to the data; 2) the items are highly discriminating; and 3) the distribution of difficulties is rectangular.

Under these circumstances, a 10-item tailored test gives as high a correlation between generated and estimated ability as a 100-item peaked conventional test. In general, Urry's data show the adaptive test to be superior to the conventional test, except for items of low discrimination. When the items are relatively imprecise, in the range of biserials of .45 which is approximately what Lord used for most of his analyses, Urry's data supports Lord's general conclusion showing a peaked conventional test to be superior for much of the population. Urry suggests that tailored tests should be considered in place of conventional tests when item-ability biserials are .65 or greater and have a relatively narrow distribution.

## Testing for Classification

All the above studies have been concerned with the problem of measurement--estimating a person's standing on a latent trait from his responses to a series of ability (or achievement) test items. Ability/achievement testing, however, is sometimes used to make classificatory decisions. Cronbach (1966) as early as 1954 and Cronbach & Gleser (1965) suggested the application of sequential or adaptive item presentation procedures to categorical decision-making. However, two studies had already applied sequential techniques to achievement testing prior to Cronbach's suggestion.

Cowden (1946) applied Wald's (1947) sequential sampling procedure in an empirical demonstration of sequential testing for assigning grades in a statistics class. Cowden's study used subsets of 20 items administered at a time, out of a pool of 200 items. Each subset was scored for each student before the next was administered; succeeding subsets were administered only when a decision could not be made with available scores. Using a set of pre-specified error tolerances, he found that decisions could be made about most students using less than one-third of the items available. Moonan (1950) applied the same sequential methods, but using real data simulation techniques to make a dichotomous (pass-fail) decision. His data showed that an average of 40 items was necessary to well approximate the decision which would be made on the basis of the parent 75 items; correlations between proportions correct on the sequential tests and the parent test were around .90. Thus, both early studies show considerable savings in terms of numbers of items required to make categorical decisions under sequential procedures.

Over twenty years later Ferguson (1971) applied the same sequential procedures to criterion-referenced achievement measurement using computer administration of achievement test items to live subjects. Ferguson's study was concerned with classifying students with respect to mastery or non-mastery at each level of a hierarchically structured achievement domain. Following the administration of each item the sequential probability ratio test was used to classify each student into one of three categories: 1) mastery; 2) non-mastery; or 3) no decision. When "no decision" occurred an additional item was administered, and the probability ratios were re-calculated. Item administration continued for each individual until a mastery or non-mastery decision was reached.

Ferguson administered his computerized sequential classification system to 75 students in grades 1 to 6 and compared the results with paper and pencil administration. Results were evaluated on several criteria. He found a 60% time savings in the computerized administration. Test-retest of the sequential procedure gave high reliability, with the reliabilities of the sequential classifications higher than those of the paper and pencil approach. Validity of the sequential approach was also found to be high.

Linn, Rock & Cleary (1970) report a real data simulation study designed to compare two sequential item administration procedures with conventional testing procedures on their effectiveness in classifying students into high and low achievement groups on the College Board's CLEP tests. Data were item responses and total scores for 4,840 students, split into development and cross-validation groups. Test items were treated in actual order of administration, and the decision rule for classification was based on log likelihood ratios. Items were "administered" to each subject until it was possible to classify him into the high or low criterion group. The sequential item administration procedure was compared to short conventional tests of from 5 to 60 items (in increments of 5), for which total test score was used to classify into achievement groups. The general conclusion derivable from Linn *et al.*'s analysis is that the sequential tests required about 50% fewer items than the conventional tests.

In general, the available classification studies using sequential procedures converge on one conclusion: sequential testing strategies can effect a considerable time savings in achievement classification. A minimum of 50% time savings in number of items administered was found in empirical studies using both paper and pencil and computer administration, as well as in two real-data simulation studies. This conclusion is further supported by Green's (1970) similar theoretical findings.

## EVALUATION

The research on adaptive testing appears to show advantages for the adaptive approaches as compared to conventional ability testing procedures. Adaptive tests show important reductions in number of items administered, with little loss of information in total scores (Bayroff & Seeley, 1967; Bryson, 1971; Cleary et al., 1968a,b; Ferguson, 1970; Krathwohl & Huyser, 1956; Linn et al., 1969, 1970); Hansen (1969) showed shorter actual testing times for computerized testing. Some adaptive testing strategies give higher validities against external criteria (Angoff & Huddleston, 1958; Hansen, 1969; Linn et al., 1969); other studies show higher correlations of adaptive test scores with underlying ability (Urry, 1970; Waters, 1964, 1971; Waters & Bayroff, 1971). For certain segments of the population, adaptive tests give considerably more precise scores, or more information per item administered (Lord, 1970, 1971a,d,e; Mussio, 1972; Paterson, 1962; Stocking, 1969); adaptive tests have been shown also to be more reliable (Angoff & Huddleston, 1958; Ferguson, 1970; Hansen, 1969). Score distributions are also affected by adaptive testing (Bayroff et al., 1960; Bayroff & Seeley, 1967; Seeley et al., 1962; Waters, 1964) with these distributions approaching equidiscriminating rectangular distributions (Hansen, 1969) and better reflecting atypical ability distributions (Paterson, 1962).

There are, of course, some negative findings concerning adaptive tests. In some studies, the expected advantages of adaptive testing were not evident from the data. In large part, however, these appear to be due to methodological difficulties of the studies themselves. Indeed, each type of study appears to have problems unique to it.

### Empirical studies

Empirical studies of adaptive testing have a number of common problems which, in some cases, have severely restricted the generalizability of their findings. These studies are, of course, limited by the characteristics of their item pool. Thus, a poorly normed item pool with low item discriminations and a poor range of item difficulties (Urry, 1970) can severely distort the findings of the empirical studies. The early studies by Bayroff & Seeley (1967) and Seeley et al. (1962), in which large numbers of testees obtain highest scores exemplify this problem. The problem is probably even more severe in the application to Bayesian adaptive procedures since they require a well-designed item pool for optimality. Yet these latter procedures are still likely to give almost optimal results in comparison to others when item pools are poorly designed, simply because they select the best item from those that do exist with maximum adaptation to individual differences among testees rather than following a pre-determined branching procedure.

Within the fixed branching models, a poorly structured branching procedure can severely vitiate the conclusions drawn from an empirical study. Bryson's (1971) study, in which items at each stage did not always progress in a meaningful order of difficulties, typifies this problem. Since the purpose of adaptive testing is to converge on an individual's ability level, a set of items structured in a way that does not follow a logical convergence procedure is unlikely to give the desired results.

The value of empirical studies is also reduced by the nature of the samples studied. In many cases the samples are simply too small to permit any general conclusions. In others, the samples represent groups of highly restricted abilities, thus limiting generalization to groups of other ability levels.

Many early empirical studies used paper and pencil administration of adaptive tests or special equipment such as punch boards. The results of these studies are, of course, confounded by the administrative complexities involved in the branched administrations. Since the adaptive tests administered in a paper and pencil or similar format require the individual to route himself through the testing procedure, additional sources of error in adaptive test scores might include the subject's willingness and his ability to follow instructions.

In spite of their limitations, however, empirical studies are an essential type of research on adaptive testing. It is only through empirical studies that the actual effects of adaptive test administration on the testee and his performance will ultimately become known. Future empirical studies of adaptive testing should be based on reasonably large numbers of subjects from carefully defined populations, using tests based on well-structured item pools normed on large and appropriate groups of subjects, with tests pre-tested to obtain appropriate kinds of score distributions and probably computer-administered to reduce extraneous sources of variance in test scores.

### Simulation studies

In the absence of well-designed empirical studies, simulation studies appear to be a valuable source of data with which to evaluate adaptive testing procedures. The "real data" simulation studies have provided important results to date and likely will continue to generate important findings. Many of these studies, however, suffer from the same limitations as the empirical studies: samples are not representative, item pools are severely restricted, and branching procedures are poorly designed, primarily because of limitations in the item pool. In addition, these studies do not include an evaluation of the possible psychological or motivational effects of adaptive testing. They can be used simply as a preliminary device for the technical comparisons of certain adaptive strategies, but results should not be considered definitive until they are replicated in empirical live testing studies.

Both Bryson (1971) and Wood (1971) compared simulation results with live administration empirical results. In both cases the simulation data gave better results than the actual computer-administered test. Bryson simulated the adaptive testing procedure on item responses of subjects who had taken a conventional test, while Wood took actual computer-administered test response patterns and used a simulated item pool. Bryson's results suggest some man-machine interaction contamination factors which affected her empirical results, while Wood's findings indicate the use of a poor item pool in his empirical study. Thus, the replication using simulation techniques of an unexpected or contradictory finding from a "real administration" empirical study can help the researcher to uncover possible design problems in his empirical study.

Monte carlo simulation studies have provided important findings concerning adaptive testing strategies. These studies eliminate as sources of error characteristics of the subjects and characteristics of the item pool. Rather, they permit the generation of item pools with known characteristics and subjects with known ability distributions. They do, however, suffer from the other problems of the "real data" simulation studies, and, due to their similarity to the theoretical studies, have the same problems inherent in those studies.

### Theoretical studies

Although theoretical studies can, in a short period of time, provide a great deal of comparative information on a variety of testing strategies, they are probably the most limited in value of any of the types of studies reported. This is not to say that they are without value--they certainly can provide some very tentative answers to specific questions. But, because of their limitations they should be carefully followed by both simulation and empirical studies to verify their conclusions.

Theoretical studies not only concern themselves with hypothetical individuals and hypothetical test items, but they must use an explicit mathematical model which might have limited relevance to what happens in actual testing. Lord (1971a) qualifies the conclusions drawn from his theoretical studies by indicating that they do not provide "fully optimal answers" to most questions of adaptive testing.

The results derived from Lord's theoretical analyses and others using similar methodologies are limited by a number of factors. First, theoretical studies must assume a specified form of the item characteristic curve for all items. These assumptions do not allow items to vary in terms of these curves. Nor have the studies to date allowed items to vary in terms of discriminations. All of Lord's analyses (and those of Mussio, 1972; Stocking, 1969) used items of fixed discrimination, a biserial correlation of .45. Urry's (1970) simulation study, however, shows that adaptive tests with higher discriminations can improve over conventional tests. While Waters & Bayroff's (1971) theoretical studies also varied item discriminations, the theoretical model forced them to keep all item discriminations equal in a given test. Urry (1970) again, using a different methodology, showed that item discriminations can vary in an adaptive test with little loss in efficiency.

In both Lord's and Bayroff's studies, number of items to be administered to an individual was fixed. The related research by Ferguson (1971) and Linn et al. (1970), as well as suggestions by Green (1970) and Weiss (1969), indicate that tailoring the number of items to be administered to a given individual might more sharply contrast adaptive and conventional testing procedures. In his analyses with guessing assumed, Lord assumes all guessing to be completely random. But Lord himself (1970), as well as Owen (1969), Urry (1970), Wood (1971) and others imply that as item difficulties get closer to the subject's ability level, the probability of random guessing decreases. Thus, in tailored testing, results derivable from a random-guessing model are not likely to be truly representative of the differential effects of tailored testing on an actual testee's test-taking behavior.

Lord's branching procedures are based simply on an individual's responses to a single test item. In this way he ignores all previous item data, thus wasting a great deal of information that can be utilized in other models, such as the Bayesian strategy (Owen, 1970; Wood, 1971) or Urry's (1970) adaptive strategy. Nor does Lord's analysis allow for the possibility of differential branching on the basis of the difficulty of incorrect answers, as implemented partially by Bayroff & Anderson (1960), thus losing some potentially valuable information in an individual's responses.

In both Lord's and Bayroff's theoretical studies, item discrimination data are based on total group data. Both conventional and adaptive tests use these same discrimination values. However, Bayroff (1969; Bayroff & Seeley, 1967) has suggested that both item difficulties and item discriminations change as a function of ability level. It is obvious that item difficulty based on a total group will not be the same as item difficulty for the same item based on a group of high ability. Item discriminations, likewise, change as a function of ability level. Data supporting this are shown by Bryson (1971). Others (e.g., Hick, 1951) imply that item discriminations for adaptive testing should be computed within an ability subgroup, rather than on total group, since as a result of all items not being administered to a total group, items need only discriminate within a specified ability level group. Therefore, a "fair" comparison of the adaptive and conventional strategies would use item discrimination data based on total group for the conventional test and item discriminations within ability groups for the adaptive test.

Following his theoretical analysis of the Robbins-Munro procedures, Lord concludes that tailored tests are, in general, technically infeasible because of the large numbers of items necessary to implement these procedures. This conclusion is, of course, derived from his analyses using the Robbins-Munro shrinking step size model. Earlier, however, Paterson (1962) showed that a different approach to a shrinking step procedure can produce significant results with small numbers of items without making the specialized assumptions involved in the Robbins-Munro process. Since Lord's theoretical analyses are based on an extremely limited set of psychometric assumptions, combined with additional very specialized mathematical assumptions, their value is only suggestive; the results of theoretical studies must be verified and extended by simulation and empirical studies in which the specialized assumptions can be relaxed and/or systematically varied.

## The Criterion Problem

The research on adaptive testing has been evaluated on the basis of a number of different criteria. In some cases, the use of different criteria in similar studies has led to somewhat different conclusions. For example, in his studies Lord concludes on the basis of information functions, that a peaked test with specified characteristics is superior to a branched test for about 70% of the ability distribution; Waters & Bayroff (1971) on the other hand, evaluate a similar pair of strategies and find that the adaptive approach has higher correlations with underlying ability. This raises the question of which of the criteria are most appropriate and which should be de-emphasized.

Correlation with paper and pencil tests. Many studies (e.g., Bryson, 1971; Linn et al., 1969) have evaluated their results in terms of the accuracy with which an adaptive test can estimate the total scores on a conventional test. If, with a given number of items, the adaptive test correlates highly with the conventional test, the results are considered to be in favor of the adaptive test. This approach, however, tends to reify the conventional test as a standard which must be met by the adaptive test. Bayroff (1964), in fact, began his work in branched testing with the hope of finding short branched tests which estimated well the scores on longer conventional tests.

The focus of adaptive testing should not be on estimating scores on a conventional test, but on improving the measurement characteristics of the scores derived from the adaptive tests. According to Lord (1971e), a good adaptive testing procedure "provides reasonably accurate measurement for examinees who would obtain near-perfect or near-zero (or near-chance-level) scores on a conventional test" (p. 228). Wood (1971) suggests that correlations with scores on conventional tests continue to perpetuate a "group testing mentality" rather than an emphasis on reducing error in estimating ability for a given individual. Rather than seeking high correlations of adaptive tests with conventional tests, an emphasis on error reduction would seek lowered correlations between the two strategies.

This latter reasoning is based partly on the findings concerning precision of measurement of adaptive vs. conventional testing strategies. The data show that both strategies give about the same errors of estimate of ability for those individuals near the center of the ability distribution. It could, therefore, be reasonably assumed that for those individuals the two procedures will correlate highly. For individuals in the extreme 30% or more of the distribution, however, conventional tests have a larger error of measurement. Scores on these tests will be highly affected by random errors, and the ordering of individuals in these areas of the ability distribution will be determined to a large part by random factors. Adaptive testing, on the other hand, maintains nearly equal precision for individuals throughout the ability range. Scores derived from adaptive tests, therefore, are more likely to be based largely on underlying ability than on random error factors.

Therefore, individuals at the extremes of ability are more likely to be ordered largely on the basis of ability.

Now, if the total score distributions for the conventional and adaptive strategies are compared, the orderings of individuals in the tails of the distributions should be different. Since product-moment correlation coefficients are means, they are affected most by changes at the extremes of the distributions--precisely where the two testing strategies are likely to order individuals differently. Thus a lowered product-moment correlation might be expected from correlating scores on conventional tests and adaptive tests, as evidence that the adaptive test is ordering individuals differently. This result might, of course, be more meaningful if compared to, say, the parallel administration of two parallel conventional tests.

Correlation with underlying ability. A number of studies (e.g., Waters & Bayroff, 1971) have correlated observed test scores with underlying ability as the criterion for evaluating adaptive tests, while Urry (1970) correlated estimated ability with generated ability in his simulation study. While this approach seems to be generally appropriate, it does have one potential problem of which future researchers in this area should be cognizant. The estimation of ability from item responses always assumes a specified mathematical model, or a set of formal assumptions. In addition, it assumes the availability of indices of item discrimination and difficulty based on that mathematical model. When adaptive test scores do not show high correlations with underlying ability, the fault may be not in the adaptive testing procedure but in the inapplicability of the model for the adaptive testing procedure. Since all testing models to date are based on assumptions derived from conventional testing, applying that model to the estimation of ability in adaptive testing might not, in some cases, be as fair a comparison as if the computations for the adaptive model were appropriate to that procedure.

Information functions. The information function can provide valuable data on the relative performance of testing strategies over a wide range of conditions. But, the use of information functions may also be limited by the inapplicability of the model to adaptive testing, since the information function utilizes computations derived from the applications of traditional test theory. Further, however, interpretation of the information functions is a highly subjective process. While Lord shows differences at or near the mean ability for adaptive and conventional testing procedures, there is no way to determine whether these differences are in any respect "significant". Lord suggests that the best 60-item adaptive test is as good as a 58-item "peaked" test. Is the difference of two items important in any respect, or do the two procedures give essentially equivalent results? Green (1970), in a re-interpretation of

Lord's data in terms of standard errors of measurement, shows that that way of looking at the results reduces the differences between the strategies even more in the middle range of abilities; at the extremes this method accentuates the precision of the adaptive approach. Thus, the researcher must interpret differences in information functions on an almost completely subjective, and highly individual, basis, thereby leading to possibly different conclusions.

Other criteria. As has been suggested, the relative utility of ability testing strategies can not be based on a single psychometric criterion, since none is wholly adequate. External validity is perhaps the ultimate criterion, but some intermediate criteria are necessary for more preliminary evaluations of various strategies.

One thus far unused but practical criterion for evaluating adaptive testing procedures might be test-retest stability data. It should be expected, because of the nearly constant precision of the branched tests, that these testing procedures would yield higher stability coefficients than would standard tests. This finding might vary with the scoring method adopted for use in branched testing, but the "best" branched testing scoring procedures should be more stable than "total scores" derived from standard tests. Stability of score estimates derived from computer-administered tests might be further improved by taking into account intra-individual adaptation patterns as reflected in item response latency information.

At the same time, other criteria are appropriate for evaluating these procedures. Such criteria include cost of test administration; costs of test scoring and reporting; time savings in test administration, both on the part of the testees and administrator; and the complexity of test administration, particularly with a view toward the effects of confounding variables. However adaptive tests are evaluated in comparison to conventional, the comparison should include a variety of criteria, both practical and psychometric, rather than a single, possibly inappropriate, criterion.

## New Problems Raised by Adaptive Testing

In its attempt to solve some of the problems inherent in conventional group paper and pencil testing, adaptive testing has raised a number of new problems waiting to be addressed by the psychometric community.

Variety of adaptive procedures. It is clear from the research already reported that there is a vast array of adaptive testing procedures which have been proposed, with an unknown number yet to be invented. While theoretical studies such as Lord's attempt to narrow down the range of possibilities, because of the limitations of those studies further efforts should be viewed with skepticism. Monte carlo studies are expensive and might not adequately reflect real testing situations, nor will "real data" simulation studies. Empirical studies are also limited, but their use is necessary if adaptive testing has any psychological effects. While there is presently no clear answer on how to narrow down the range of adaptive testing strategies, the most fruitful approach might be the development and implementation of test theory specifically designed for use in adaptive testing.

Scoring methods. Various scoring rules have been proposed for adaptive testing. Scores include number correct, final difficulty score, average difficulty score, and difficulty of the n+1th item, as well as various approaches to scoring two-stage models. A major problem also common to conventional tests is that all testees with a given final score have not necessarily gotten the same items correct. However, in adaptive testing the problem is more critical because the variety of items available is greater. The end result of the problem may be difficulties in interpreting scores, which could lead to legal problems in the use of test scores (Lord, 1971b). While Lord's flexilevel test avoids this problem, since all people who get a given score have taken the same items, the same simplicity in interpretation is not available in most other methods of adaptive testing. Explaining a test score to laymen will be especially difficult in approaches such as the Bayesian strategies, which assume a rather complex model underlying test scores. Beyond that, however, the optimal method of scoring adaptive tests from a psychometric viewpoint will remain an important issue for some time.

Appropriateness of methods of item analysis. The possible inappropriateness of classical test theory for adaptive testing is reflected in the inappropriateness of test construction methods for problems of adaptive testing. In particular, the question of the appropriateness of the methods of computing item discrimination indices and item difficulty indices can be questioned.

Traditional methods of determining item discrimination are based on variations of the biserial correlation of item responses with total score or with score on the latent ability continuum. In these computations for conventional testing all subjects are assumed to have responded to each test item; hence the biserial computation is based on the data for the entire group of subjects, who vary across the ability continuum. In essence, the biserial correlation reflects the mean difference between all subjects who correctly answer an item and all subjects whose response to that item is incorrect.

The appropriateness of these computations has been implicitly or explicitly questioned by a number of studies in adaptive testing. Paterson (1962) suggests that items to be used in adaptive testing can have low discriminations as computed on a total group, since they do not have to discriminate across a range of abilities, yet they can be useful in adaptive testing since they can discriminate within a narrow ability range at some point on the ability continuum. Bryson (1971) suggests that the discriminating power of an item to be used in an adaptive test be based on the point-biserial correlation of item response and total score for the subjects who take that item. Thus, the discrimination index would be computed on a group more homogeneous with respect to ability; the discrimination then is between those who answer an item correctly and those who answer it incorrectly, _within a limited ability range_. This approach to item analysis was implemented by Bryson (1971) and Cleary _et al_. (1968a). Bryson's (1972) data show this method of item analysis to produce highest validities for very short tests as compared to more traditional methods of item selection.

The applicability of traditional indices of item difficulty can also be questioned. Hick (1951) suggests that the appropriate test item to be administered to an individual is an item of 50% difficulty for individuals of a given estimated ability, since that item provides the most information in its responses. This suggestion is echoed by Levitt (in Harman, Helm & Loye, 1968), who likens ability measurement to the problems of estimating points on a psychometric function, and by Lord (1972) and Novick (1969). Since a given item with probability of .50 for a group of specified ability might not be identified by standard item analysis procedures as an appropriate test item for administration under adaptive testing, the construction of adaptive tests using standard item difficulties should be carefully examined.

_Effects of chance_. The effects of chance success on multiple choice test items needs to be carefully considered in the construction and administration of adaptive ability tests. Bayroff (1969) has suggested that chance responding to items in a multi-stage branched test may have a greater effect on test scores than in conventional tests. His reasoning appears to be based on the much smaller number of items used in multi-stage branched tests as compared to conventional tests. He suggests that a few items in a row correctly answered by chance might lead an individual down an inappropriate path in the branched strategy, and that there may not be sufficient succeeding items in a short branched test to allow "recovery" to an appropriate ability level for that individual. It should be noted, however, that this criticism applies only to the multi-stage pyramidal strategy and only to the case where the termination rule does not use some explicit convergence criterion, allowing number of items administered to vary for each individual.

The differential effects of chance in adaptive testing as compared to conventional tests might be viewed in a contrasting way. In the typical "peaked" conventional test in which items are concentrated around some average value, only individuals whose abilities lie at the average value will take test items which are of appropriate difficulty for them. For all individuals above the ability level of the test items, the items will be too easy and guessing, and therefore chance successes, will not likely occur. It is only the individuals of ability below the average ability of the peaked test who are likely to guess, with the probability of guessing--and therefore, chance success--increasing with decreasing ability. In the typical non-peaked test, all individuals except for those of highest ability will be presented with some test items which are above their ability level. Thus, chance successes are possible for most testees. Explicit models of guessing behavior which take account of these hypotheses have been proposed by Urry (1970) and Wood (1971).

The purpose of adaptive testing, however, is to keep test items at a level of difficulty appropriate to a given individual. Thus, the adaptive procedure searches for the ability level of the testee and presents test items as close to that level as possible, since it is these items which yield maximum information (Hick, 1951). Since adaptive procedures tend to minimize the number of items which are too difficult for a given individual, they should also tend to reduce guessing and therefore the probability of chance successes. Bayroff (1964, 1969) suggests that keeping test items at a level relevant to an individual's ability might reduce carelessness errors; both Green (1970) and Lord (1970) suggest similar motivational effects by adjusting difficulties to an individual's ability level. Hansen (1969) has shown that decreases in guessing do occur when difficulties are tailored to an individual's ability level. A relevant topic for future research in adaptive testing, then, is to further specify the exact effects of guessing on tailoring test items to each individual's ability level during the testing process.

<u>Termination rules</u>. In conventional testing two termination rules are essentially universal: 1) every individual takes every item in the test; or 2) everyone

terminates at the end of a specified period of time, regardless of the number of items completed. Adaptive testing, however, permits the development of a number of new rules for termination of testing. While many writers on adaptive testing (e.g., Bayroff _et al_., 1960, 1967; Cleary _et al_., 1968a,b,; Lord, 1970, 1971a,d,e) have studied adaptive testing using a fixed number of items for all individuals, that procedure appears to ignore the likelihood of individual differences in convergence, thus vitiating a prime element of the adaptive capabilities of the testing procedures.

A number of writers have suggested adaptive termination rules. Novick (1969), Urry (1970) and Wood (1971) continue testing until the error in the ability estimate converges on some pre-specified value; this approach has also been suggested by Green (1970) and Weiss (1969). Lord (1972) and others have suggested that testing continue until a level of difficulty is reached at which the individual gets 50% of the items correct and 50% incorrect; since that level provides the most information about an individual's ability, it can be assigned as his final ability level.

While considerable research needs to be done in developing and validating termination rules for adaptive testing, some rules, such as the latter one proposed, can be questioned on purely logical grounds. It would seem more logical, in the case of a multiple choice response, to terminate testing at the highest difficulty level at which an individual gets more than 1/n items correct, where n is the number of response choices in each test item; that is, to identify as his final ability score the highest difficulty level at which he responds correctly beyond a chance level. Only in the case of true-false or other dichotomous response test items would this termination rule agree with Lord's suggestion.

_Information utilization_. In recent years, a number of psychometricians (e.g., Echternacht, 1972; Shuford, Albert & Massengill, 1966; Wang & Stanley, 1970) have suggested differential response option weighting or response-determined scoring as means of improving the reliability and validity of ability tests by making greater use of information provided in incorrect answers to multiple choice test items. Research in this area shows some promise for these approaches (e.g., Coombs, Milholland & Womer, 1956; Davis & Fifer, 1959; Feldman & Markwalder, 1971), although all findings are not yet consistently in favor of the approach.

Adaptive testing permits the extension of differential response option weighting to differential response option _branching_. In this procedure, the choice of the next item to be administered following an incorrect response is made on the basis of the "incorrectness" of the response given. Thus, a person who chose a response option frequently chosen by persons of low average ability would be branched to a much easier next item than the individual who chose an incorrect option chosen by persons of higher average ability. Such an approach would use all the information available in a subject's response record, perhaps permitting quicker convergence on the appropriate ability level for each testee and possibly capitalizing better on non-chance guessing among incorrect response alternatives. Such a procedure has been suggested by Wood (1971) and used by Bayroff _et al_. (1960) on the first item only of his multi-stage branching model. Considerable empirical research remains to be done, however, to systematically investigate the utility of this approach.

## Implementing Adaptive Testing

_Paper and pencil tests_. Since paper and pencil testing has dominated in the implementation of ability testing for over 50 years, it is natural to attempt to capture the advantages of adaptive testing within a paper and pencil format. Early research with adaptive ability measurement (Bayroff _et al_., 1960; Seeley _et al_., 1962; Wood, 1969) studied multi-stage pyramidal tests administered by paper and pencil. These tests involved the use of complicated instructions to the testee or answer sheets which informed the testee of the correctness of his response to each item, so that appropriate branching could occur.

Bayroff _et al_. (1960) found it necessary to include in their branched paper and pencil tests a number of "buffer" items so that the correct branching sequence would be followed by each testee. Scoring of these branched tests was simple in that the score was the difficulty level of the last item reached by a given testee. However, this score was valid only if the testee had followed the branching instructions. Thus, scoring of the paper and pencil branched test was considerably more complex than the conventional paper and pencil test since the response path of each testee had to be individually validated by the scorer. Seeley _et al_.'s (1962) data implementing Bayroff's test showed that the paper and pencil branched tests were more time-consuming to construct, took longer to administer, and posed difficult problems in scoring due to verification of routing, in comparison to conventional tests. Their data also showed a substantial number of testees not following the branching instructions, thus invalidating their test records. This latter finding also appeared in Wood's (1969) data using a paper and pencil branched test.

Despite the negative data available concerning the implementation of adaptive testing in a paper and pencil medium, the idea still appears to be alive; Lord (1971b) recently proposed his flexilevel test for paper and pencil administration. The testing procedure requires that the answer sheet inform the testee of the correctness of his response and that he proceed to different items depending on whether his response is correct or incorrect. No empirical data are as yet available on the problems involved in administering flexilevel tests, but it would appear that only a highly motivated and reasonably capable testee would produce a valid response record from paper and pencil administration of a flexilevel test.

Testing machines. Because of the difficulties involved in administering adaptive tests by paper and pencil methods, testing machines have been proposed as a logical alternative. Typical of such proposals is Bayroff's (1964) testing machine. This device was built around a 35 mm. slide projector and was capable of administering linear, two-stage, and pyramidal tests. His machine also recorded response latency information, had the capability of stopping testing if the testee's score fell above or below pre-specified cutting points, and allowed the examinee to choose one or more "tentative" answers before recording the "final" answer to an item. For a variety of reasons, however, Bayroff's testing machine was never built.

More recently, Elwood & Griffin (1972) report on the successful development and application of a more complex testing machine, although its use is currently devoted to administration of the Wechsler Adult Intelligence Scale (WAIS). This machine does no branching; rather, it simulates administration of the WAIS as if it were administered by a human examiner. The purpose is to eliminate examiner variables, not to adapt the test to individual differences. Elwood & Griffin's results show that such automated administration does, for the most part, yield scores which are comparable to those obtained by human examiners. Machine administration of the digit span test, however, was not comparable to that of human administration. Thus, in some cases testing machines can change the nature of the variable being measured. Whether the changes are toward greater reliability and validity of measurement remains to be seen. A further problem with the WAIS testing machine is the large amount of "set-up" time required to prepare the machine for administration of the test to subsequent testees.

Computer administration. The advent and growth of timeshared computer facilities has great promise for the implementation of adaptive ability testing. Computer control of adaptive test administration completely avoids the problems inherent in paper and pencil adaptive testing and in the use of some testing machines. When the computer controls branching, the branching decisions are completely out of the testee's hands. The computer presents a test item, records the response, branches in almost an infinite number of ways to the next test item, and presents the selected item to the testee. Under computer administration, invalid response sequences will not occur; thus, every testee will produce a valid branching record. Furthermore, computer administration will not require the examinee to be highly motivated or capable of following instructions about branching; the examinee's participation is passive, with his attention directed solely to the solution of the test questions, once he has learned how to operate the testing terminal. With the exception of physical (as opposed to symbolic) stimuli, reconstruction of a stimulus which is altered in the process of administration is an instantaneous process for the computer, not requiring additional administrator intervention.

A variety of other advantages have been proposed for administration of psychological tests under computer control. Among these advantages Cronbach (1970, p. 73) includes excellence of standardization, precision of timing, release of testers for other duties, the computer's infinite "patience," control of bias and reduction of test anxiety, and the integration of testing and learning. Stillman, Roch, Colby, & Rosenbaum (1965), in applying computer methods to the administration of personality items, suggest a "neutrality" effect, reducing examiner effects which affect test performance. Hansen, Hedl, & O'Neill (1971) support this idea in the context of achievement testing by suggesting that computer administration of achievement tests will be more neutral than administration of the same test items by teachers. Such neutrality, they suggest, by eliminating biases due to the "dyadic interaction" of student and teacher, may lead to increases in both reliability and validigy; Hedl (1971) suggests a similar reduction in bias, leading him to develop non-branched computer administration of an individual intelligence test. Johnson (1967) discusses data which show less variability in task performance under conditions of computer (vs. experimenter) administration. He reasons that the reduced variability might be due to reductions in error variance; as a result, computer administration may be more sensitive to "real" effects than other modes of stimulus administration.

In addition to possibly reducing error variance, computer administration of ability tests opens a host of new approaches to ability measurement. Morrison (in Harman *et al.*,

1968) suggests that computerized ability testing would allow the measurement of both new content and modes of abilities. A start in this direction has been reported by Cory (1972). Cory's research concerns the development of ability tests administered by computer to measure a variety of perceptual abilities not easily measurable by paper and pencil techniques. The tests include tests of object, number and word memory, each using controlled exposure times, perceptual speed and closure, and movement detection and memory for patterns. Two additional tests are put in the format of games to measure specific kinds of verbal reasoning abilities. The game format was chosen in an attempt to motivate testees of low and marginal ability to perform to their maximum on the tests.

Green (1970), Holtzman (1970) and Hubbard (1966) have suggested that computerized test administration can be used to study an individual's problem solving abilities. This approach would represent a within-problem branching sequence in which a series of interdependent questions are organized into a problem-oriented structure; the testee's path through the structure would serve as an indication of his ability to reason in specified ways. Newell's (in Harman _et al_., 1968) suggestion of using the computer to study "coping strategies" is closely related to this application.

Computer administration of ability tests also makes feasible the use of confidence weighting techniques (Shuford _et al_., 1966) for ability test items. Closely related to this is the suggestion by Green (1970) and Holtzman (1970) that the testee be permitted to continue answering until he gets each item correct; the sequence of responses chosen then becomes additional information usable in deriving individual test scores. This approach was used (but not explicitly studied) by Wood (1971); a recent study using a paper and pencil variation of this scoring method (Gilman & Ferry, 1972) shows higher reliability for scores derived from this type of test response procedure.

The use of item response latency data is an additional benefit derivable from computerized test administration. Response latencies might be usable in conjunction with confidence weighting procedures. Green (1970) suggests that a careful analysis of latency data could lead to the identification of guessing behavior on specific test items. Should guessing be identifiable in this way, guessed responses could be eliminated from a testee's score, thus possibly reducing error variance. The measurement of response latencies also has implications for theories of ability measurement, since it will assist in differentiating among those individuals who respond correctly on a given test item in terms of speed of response, thus distinguishing the "fast but correct" testee from the "slow but correct" responder.

Immediate knowledge of results is another potential benefit of computerized ability testing. Bayroff (1964) and Ferguson & Hsu (1971), among others, have suggested that immediate feedback to testees on their performance on each test item might have positive motivating effects, with subsequent positive benefits in more reliable or valid test scores. This potential positive effect of immediate feedback is even more likely when the testing strategy is programmed to provide large proportions of positive feedback to the testee. The effect of such positive feedback in the testing situation might be more prominent among members of minority groups for whom testing situations are likely to carry more negative than positive affect. Through appropriate computerized testing it might be possible to transform the testing situation into a positive experience, increasing test-taking motivation and reducing test-taking anxiety.

Computer administration of adaptive tests could permit control of the degree of precision attached to any given individual's test score (Norman, in Harman et al., 1968; Weiss, 1969). The computer can calculate after each test item administered some kind of "standard error of measurement" term to be attached to the ability estimate. Further, since increasing the number of items administered in an adaptive fashion will, in general, decrease the error estimate, tailoring the number of items administered to each individual (Ferguson, 1971; Green, 1970), based on sequential error estimates, will in effect permit the tester to control the degree of precision attached to the obtained test score. Cronbach (1966) suggests a similar procedure in a decision context, in which the number of observations for each individual is tailored to specified error rates in the decision function. In ability testing, this approach is currently operationalized only in the Bayesian models (Novick, 1969; Owen, 1969, 1970; Wood, 1971).

## CONCLUSIONS

Research available on adaptive testing shows considerable promise for the superiority of these methods over conventional ability testing procedures. Using a variety of research approaches and a number of different criteria, adaptive tests have been shown to be: 1) considerably shorter than conventional tests, with little or no loss in validity or reliability; 2) more reliable than conventional tests in several studies and yielding more nearly constant precision than standard tests throughout the range of abilities; and 3) in several cases more valid, as measured against an external criterion, than are conventional tests. Adaptive tests also have promise of being more "fair" to minority group members in that the range of item difficulties is less likely to result in frustrating or negative experiences, thus permitting ability estimates less confounded by error.

Although applications of adaptive tests raise many new problems in psychometric research, their future development as an important approach to ability measurement seems assured by their potential value. Because of the complexity of some of the branching decisions which need to be made in adaptive testing, neither paper and pencil methods of administration nor special testing machines will allow all the future benefits of adaptive testing to surface. Full utilization of the capabilities of adaptive testing will be realized only through the use of time-shared computer systems as test administration devices. Such computerized test administration will permit the development of new methods of ability testing and new theoretical approaches, leading to what Green (1970, p. 194) calls "the inevitable computer conquest of testing."

## References

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

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

Baratz, S. S. Effect of race of experimenter, instructions, and comparison population upon level of reported anxiety in Negro subjects. *Journal of Personality and Social Psychology*, 1967, *7*, 194-196.

Bayroff, A. G. Feasibility of a programmed testing machine. U. S. Army Personnel Research Office Research Study 64-3, November 1964.

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

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

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

Berger, V. F., Munz, D. C., Smouse, A. D. & Angelino, H. The effects of item difficulty sequencing and anxiety reaction type on aptitude test performance. *Journal of Psychology*, 1969, *71*, 253-258.

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

Boldt, R. F. Study of linearity and homoscedasticity of test scores in the chance range. *Educational and Psychological Measurement*, 1968, *28*, 47-60.

Brenner, M. H. Test difficulty, reliability, and discrimination as functions of item difficulty order. Journal of Applied Psychology, 1964, 48, 98-100.

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

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.

Caldwell, M. B. & Knight, D. The effect of Negro and White examiners on Negro intelligence test performance. Journal of Negro Education, 1970, 39, 177-179.

Cieutat, V. J. Examiner differences with the Stanford-Binet IQ. Perceptual and Motor Skills, 1965, 20, 317-318.

Clark, C. A. The use of separate answer sheets in testing slow-learning pupils. Journal of Educational Measurement, 1968, 5, 61-64.

Cleary, T. A., Linn, R. L. & Rock, D. A. An exploratory study of programmed tests. Educational and Psychological Measurement, 1968, 28, 345-360. (a)

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

Cohen, E. Is there examiner bias on the W-B? Proceedings of the Oklahoma Academy of Science, 1950, 31, 150-153.

Coombs, C. H., Millholland, J. E. & Womer, F. B. The assessment of partial knowledge. Educational and Psychological Measurement, 1956, 16, 13-37.

Cory, C. H. First year's progress report: A job element approach to the validation of perceptual measures. June 1972 (unpublished).

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

Cronbach, L. J. Essentials of psychological testing. (3rd ed.), New York: Harper and Row, 1970.

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. & Gleser, G. C. Psychological tests and personnel decisions. Urbana: University of Illinois Press, 1965.

Davis, F. B. Item analysis in relation to educational and psychological testing. Psychological Bulletin, 1952, 49, 97-121.

Davis, F. B. & Fifer, G. The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. Educational and Psychological Measurement, 1959, 19, 159-170.

Donahue, D. & Sattler, J. M. Personality variables affecting WAIS scores. Journal of Consulting and Clinical Psychology, 1971, 36, 441.

DuBois, H. A history of psychological testing. Boston: Allyn and Bacon, 1970.

Ebel, R. L. Expected reliability as a function of choices per item. Educational and Psychological Measurement, 1969, 29, 565-570.

Echternacht, G. F. Theuse of confidence testing in objective tests. Review of Educational Research, 1972, 42, 217-236.

Egeland, B. Examiner expectancy: effects on the scoring of the WISC. Psychology in the Schools, 1969, 6, 313-315.

Elwood, D. L. & Griffin, H. R. Individual intelligence testing without the examiner: reliability of an automated method. Journal of Consulting and Clinical Psychology, 1972, 38, 9-14.

Exner, J. E. Jr. Variations in WISC performance as influenced by differences in pre-test rapport. Journal of General Psychology, 1966, 74, 299-306.

Feldman, D. H. & Markwalder, W. Systematic scoring of ranked distractors for the assessment of Piagetian reasoning levels. Educational and Psychological Measurement, 1971, 31, 347-362.

Ferguson, R. L. A model for computer-assisted criterion-referenced measurement. Paper presented at the National Council on Measurement in Education meetings, March 1970, Minneapolis.

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

Ferguson, R. L. & Hsu, T. The application of item generators for individualizing mathematics testing and instruction. Report 1971/14, University of Pittsburgh Learning Research and Development Center, 1971.

Flaugher, R. L., Melton, R. S. & Myers, C. T. Item rearrangement under typical test conditions. *Educational and Psychological Measurement*, 1968, *28*, 813-824.

Forrester, B. J. & Klaus, R. A. The effect of race of the examiner on intelligence test scores of Negro kindergarten children. *Peabody Papers in Human Development*, 1964, *2*, 1-7.

Frandsen, A. N., McCullough, B. R. & Stone, D. R. Serial versus consecutive order administration of the Stanford-Binet Intelligence Scales. *Journal of Consulting Psychology*, 1950, *14*, 316-320.

Frary, R. B. & Zimmerman, D. W. Effect of variation in probability of guessing correctly on reliability of multiple-choice tests. *Educational and Psychological Measurement*, 1970, *30*, 595-605.

Gilman, D. A. & Ferry, P. Increasing test reliability through self-scoring procedures. *Journal of Educational Measurement*, 1972, *9*, 205-207.

Gordon, L. V. Right-handed answer sheets and left-handed testees. *Educational and Psychological Measurement*, 1958, *18*, 783-785.

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

Greenwood, D. I. & Taylor, C. Adaptive testing in an older population. *Journal of Psychology*, 1965, *60*, 193-198.

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.

Hansen, D. N., Hedl, J. J. Jr. & O'Neill, H. F. Jr. Review of automated testing. Technical Memo No. 30, Computer Assisted Instruction Center, Florida State University, 1971.

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

Hata, Y., Tsudzuki, A., Kuze, T. & Emi, Y. Relationships between the tester and the subject as a factor influencing on the intelligence test score: I. *Japanese Journal of Psychology*, 1958, *29*, 95-99.

Hayward, P. A comparison of test performance on 3 answer sheet formats. *Educational and Psychological Measurement*, 1967, *27*, 997-1004.

Hedl, J. J. Jr. An evaluation of a computer-based intelligence test. Technical Report No. 21, Computer Assisted Instruction Center, Florida State University, 1971.

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 and Row, 1970.

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.

Huck, S. W. & Bowers, N. D. Item difficulty level and sequence effects in multiple-choice achievement tests. *Journal of Educational Measurement*, 1972, *9*, 105-111.

Hutt, M. L. A clinical study of "consecutive" and "adaptive" testing with the revised Stanford-Binet. *Journal of Consulting Psychology*, 1947, *11*, 93-103.

Johnson, E. S. The computer as experimenter. *Behavioral Science*, 1967, *12*, 484-489.

Katz, I. & Greenbaum, C. Effects of anxiety, threat, and racial environment on task performance of Negro college students. *Journal of Abnormal and Social Psychology*, 1963, *66*, 562-567.

Katz, I., Roberts, S. O. & Robinson, J. M. Effects of task difficulty, race of administrator, and instructions on digit-symbol performance of Negroes. Journal of Personality and Social Psychology, 1965, 2, 53-59.

Klosner, N. C. & Gellman, E. K. The effect of item arrangement on classroom test performance. Paper presented at the meeting of the Eastern Psychological Association, April 1971.

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

LaCrosse, J. E. Examiner reliability on the Stanford-Binet Intelligence Scale (Form L-M) in a design employing White and Negro examiners and subjects. Unpublished Masters thesis, University of North Carolina, 1964.

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. The development and evaluation of several programmed testing methods. Educational and Psychological Measurement, 1969, 29, 129-146.

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).

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

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 and Row, 1970.

Lord, F. M. Robtins-Munro procedures for tailored testing. *Educational and Psychological Measurement*, 1971, *31*, 3-31. (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. & Novick, M. R. *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1968.

MacNicol, K. Effects of varying order of item difficulty in an unspeeded verbal test. Unpublished manuscript, Princeton, N. J., Educational Testing Service, 1956.

Mandler, G. & Sarason, S. B. A study of anxiety and learning. *Journal of Abnormal and Social Psychology*, 1952, *47*, 166-173.

Marine, E. L. The effect of familiarity with the examiner upon Stanford-Binet test performance. *Teachers College Contributions to Education*, Columbia University, 1929, No. 381.

Marso, R. N. Test item arrangement, testing time, and performance. *Journal of Educational Measurement*, 1970, *7*, 113-118.

Masling, J. The effects of warm and cold interaction on the administration and scoring of an intelligence test. *Journal of Consulting Psychology*, 1959, *23*, 336-241.

Matarazzo, J. D., Ulett, G. A., Guze, S. B. & Saslow, G. The relationship between anxiety level and several measures of intelligence. *Journal of Consulting Psychology*, 1954, *18*, 201-205.

Merwin, J. C. New Measurement Research Center answer sheets and Differential Aptitude Test norms. Student Counseling Bureau Newsletter. Minneapolis: Office of the Dean of Students, University of Minnesota, April, 1963.

Miller, J. O. & Phillips, J. A. A preliminary evaluation of the Head Start and other metropolitan Nashville kindergartens. Unpublished manuscript, Demonstration and Research Center for Early Education, George Peabody College for Teachers, Nashville, 1966.

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.

Morris, L. W. & Liebert, R. M. Effects of anxiety on timed and untimed intelligence tests: another look. *Journal of Consulting and Clinical Psychology*, 1969, *33*, 240-244.

Munz, D. C. & Smouse, A. D. The interaction effects of item difficulty sequence and achievement anxiety reaction on academic performance. *Journal of Educational Psychology*, 1968, *59*, 370-374.

Murdy, W. G. Jr. The effect of positive and negative administrations on intelligence test performance. Dissertation Abstracts, 1962, *23*, 1076.

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

Nichols, R. C. The effect of ego involvement and success experience on intelligence test results. *Journal of Consulting Psychology*, 1959, *23*, 92.

Nitardy, J. R., Peterson, C. D. & Weiss, D. J. Differential influence of test format variables on ability test performance. *Proceedings of the 77th Annual Convention of the American Psychological Association*, 1969, 139-140.

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

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

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

Peters, D. L. & Messier, V. The effects of question sequence upon objective test performance. *Alberta Journal of Educational Research*, 1970, *16*, 253-265.

Plumb, G. R. & Charles, D. C. Scoring difficulty of Wechsler Comprehension responses. *Journal of Educational Psychology*, 1955, *46*, 179-183.

Quereshi, M. Y. Intelligence test scores as a function of sex of experimenter and sex of subject. *Journal of Psychology*, 1968, *69*, 277-284.

Rasch, G. An individualistic approach to item analysis. In P. F. Lazarsfeld & N. W. Henry (Eds.), *Readings in mathematical social science*. Chicago: Science Research Associates, 1966. (a)

Rasch, G. An item analysis that takes individual differences into account. *British Journal of Mathematical and Statistical Psychology*, 1966, *19*, 49-57. (b)

Sacks, E. L. Intelligence scores as a function of experimentally established social relationships between child and examiner. *Journal of Abnormal and Social Psychology*, 1952, *47*, 354-358.

Sarason, S. B., Mandler, G. & Craighill, P. G. The effect of differential instructions on anxiety and learning. *Journal of Abnormal and Social Psychology*, 1952, *47*, 561-565.

Sattler, J. M. Statistical reanalysis of Canady's "The effect of 'rapport' on the IQ: a new approach to the problem of racial psychology." *Psychological Reports*, 1966, *19*, 1203-1206.

Sattler, J. M., Hillix, W. A. & Neher, L. A. Halo effect in examiner scoring of intelligence test reponses. *Journal of Consulting & Clinical Psychology*, 1970, *34*, 172-176.

Sattler, J. M. & Winget, B. M. Intelligence testing procedures as affected by expectancy and IQ. *Journal of Clinical Psychology*, 1970, *26*, 446-448.

Sax, G. & Carr, A. An investigation of response sets on altered parallel forms. *Educational and Psychological Measurement*, 1962, *22*, 371-376.

Sax, G. & Cromack, T. The effects of various forms of item arrangements on test performance. *Journal of Educational Measurement*, 1966, *3*, 309-311.

Schwartz, M. L. The scoring of WAIS comprehension responses by experienced and inexperienced judges. *Journal of Clinical Psychology*, 1966, *22*, 425-427.

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.

Siegman, A. W. The effect of manifest anxiety on a concept formation task, a non-directed learning task, and on timed and untimed intelligence tests. Journal of Consulting Psychology, 1956, 20, 176-178.

Simon, W. E. Expectancy effects in the scoring of vocabulary items: a study of scorer bias. Journal of Educational Measurement, 1969, 6, 159-164.

Smith, H. W. & May, W. T. Influence of the examiner on the ITPA scores of Negro children. Psychological Reports, 1967, 20, 499-502.

Smouse, A. D. & Munz, D. C. The effects of anxiety and item difficulty sequence on achievement testing scores. Journal of Psychology, 1968, 68, 181-184.

Spache, G. Serial testing with the revised Stanford-Binet Scale, Form L, in the test range II-XIV. American Journal of Orthopsychiatry, 1942, 12, 81-86.

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

Stevenson, H. W. & Allen, S. Adult performance as a function of sex of experimenter and sex of subject. Journal of Abnormal and Social Psychology, 1964, 68, 214-216.

Stillman, R., Roth, W. T., Colby, K. M. & Rosenbaum, C. P. An on-line computer system for initial psychiatric inventory. American Journal of Psychiatry, 1965, 125 (No. 7 supplement), p. 8-11.

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. Boston: Houghton Mifflin, 1960.

Thorndike, R. L. Reliability. In E. F. Lindquist (Ed.), Educational Measurement. Washington, D. C.: American Council on Education , 1951.

Tsudzuki, A., Hata, Y. & Kuze, T. A study of rapport between examiner and subject. Japanese Journal of Psychology, 1956, 27, 22-28.

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

Wald, A. Sequential analysis, New York: Wiley, 1947.

Walker, R. E., Hunt, W. A. & Schwartz, M. L. The difficulty of WAIS comprehension scoring. Journal of Clinical Psychology, 1965, 21, 427-429.

Wang, M. W. & Stanley, J. C. Differential weighting: a review of methods and empirical studies. Review of Educational Research, 1970, 40, 663-705.

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

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

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

Wechsler, D. Manual for the Wechsler Adult Intelligence Scale. New York: Psychological Corporation, 1955.

Weiss, D. J. Individualized assessment of differential abilities. Paper presented at the 77th annual convention of the American Psychological Association, Division 5, September 1969.

Whitcomb, M. A. The IBM answer sheet as a major source of variance on highly speeded tests. Educational and Psychological Measurement, 1958, 18, 757-759.

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. Fully adaptive sequential testing: a Bayesian procedure for efficient ability measurement. Unpublished manuscript, 1972.

Young, D. K. Digit span as a function of the personality of the experimenter. American Psychologist, 1959, 14, 375.

# Comparison of Four Empirical Differential Item Scoring Procedures

Isaac I. Bejar

and

David J. Weiss

Research Report 73-2

Psychometric Methods Program
Department of Psychology
University of Minnesota

Paper presented at the 81st Annual Convention
of the
American Psychological Association
Division 5

August 1973

# ABSTRACT

The effect of four empirical option weighting scoring procedures on validity and reliability was investigated by means of simulated tests of varying degrees of inter-item correlation. It was found that the effect of empirical option weighting on validity and reliability depends upon inter-item correlation. From a practical point of view, it was concluded that scoring tests with other than 0-1 weights results in increases in reliability and validity that are negligible within the range of inter-item correlation encountered in typical published tests.

# Comparison of Four Empirical Differential Item Scoring Procedures[1]

Isaac I. Bejar and David J. Weiss
University of Minnesota

Several scoring procedures, including differential option weighting scoring (DOWS) have been proposed to improve the psychometric characteristics of multiple-choice tests. DOWS is a procedure by which each option within an item is assigned a weight that corresponds to the merit of that alternative. Weights are assigned so that the correct alternative receives the highest weight while the poorest option receives the smallest weight. A subject's total score is the sum of the weights of the options chosen by him. Conventional scoring, in contrast, assigns a weight of 1 to the keyed or correct option and 0 to all others. The total score under conventional scoring is, of course, the number of correct answers.

DOWS procedures are based on the contention that reliability and validity should increase as a result of an increase in the proportion of reliable variance due to the finer discrimination afforded by the DOWS scoring procedures. Research on DOWS procedures (Davis and Fifer, 1959; Hendrickson and Green, 1972; Reilly and Jackson, 1972; Sabers and White, 1969; reviewed by Wang and Stanley, 1970) shows that empirical DOWS techniques do increase the reliability of a set of scores. However, increases in validity have not been reported. A number of explanations have been advanced for the lack of increase in validity (e.g., Hendrickson, 1970; Reilly

[1] The assistance of the University of Minnesota Computer Center in providing computer time is gratefully acknowledged.

and Jackson, 1972). This paper explores the possibility that variations in the homogeneity of the tests affects the reliability and validity of tests under DOWS procedures. In order to control and manipulate the homogeneity of tests, the present study used artificially generated tests which were "administered" to groups of hypothetical subjects.

## METHOD

Generation of the Data. A response data matrix was generated by means of an extension to the polychotomous case of a simulation model proposed by Shoemaker and Osburn (1970) for dichotomous items:

$$P_{ijk} = \Phi[(C_i - A_{jk})B_j] - \Phi[(C_i - A_{jk-1})B_j] \qquad (1)$$

where:

$P_{ijk}$ = probability of subject i choosing option k in item j

$C_i$ = position of subject i on the underlying trait (i.e., ability)

$A_{jk}$ = difficulty parameter of option k in item j

$B_j$ = discrimination parameter of item j

$\Phi$ = standard normal c.d.f.

The simulation of a data response matrix involved the following steps: 1) "Subject ability" was randomly sampled from a population N(0,1), within ±3 standard deviations; 2) Items were sampled from a population of items rectangularly distributed in difficulties ranging from .27 to .73 (in traditional terminology). The "difficulty" (i.e., proportion choosing each of the remaining alternatives) was set to a constant = $(1-P_{j1})/4$. Each item had five alternatives ordered so that alternative 1 was the best or keyed alternative and alternative 5 was the poorest; 3) Item discriminations were constant for each of the 40 items within a test so that each test was unidimensional with inter-item tetrachoric correlations of .09, .16, .25, .36, .49, and .81 respectively; 4) values of C, A and B were substituted into (1) to determine for each subject the probability of choosing each of the alternatives. The option actually "chosen" by a given "subject" was determined by sampling a random number from a rectangular distribution with range 0 to 1. The cumulative response probability (beginning from the best alternative) was then compared to the random number. The alternative chosen was the one at which the cumulative probability exceeded the random number. No guessing was assumed.

Scoring Prodedures. The following scoring schemes were applied to each of the six response data matrices: conventional scoring (0-1), biserial weights (BIS), point biserial weights (PBIS), reciprocal averages weights (RAV) and theoretical weights (TW). BIS weights were based on the biserial correlation of each alternative with total score while PBIS were based on the point biserial correlations. RAV weights were the mean total score of subjects choosing each alternative. Theoretical weights were the estimated difficulty parameter ($A_{jk}$'s) of the response model (1). BIS, PBIS and RAV weights were iterated until the increase in Hoyt reliability was less than .05. The weights for each test were developed on 100 "subjects" and cross-validated on an independently generated sample of 100 "subjects."

Evaluation Criteria. Hoyt reliabilities and product-moment correlations between total score and underlying trait ("validities") were computed for each scoring scheme.

## RESULTS

Table 1 reports the cross-validated reliability and validity coefficients. Contrary to what might have been expected from earlier

Table 1

Cross-validated Reliability and Validity Coefficients for O-1 Scoring and Four DOWS Procedures as a Function of Inter-Item Correlation

| | Inter-item correlation ($r_{tet}$) | | | | | |
|---|---|---|---|---|---|---|
| Scoring method | .09 | .16 | .25 | .36 | .49 | .81 |
| Hoyt reliability | | | | | | |
| O-1 | 725 | 813 | 869 | 901 | 948 | 978 |
| TW | 621 | 798 | 873 | 912 | 957 | 988 |
| BIS | 685 | 798 | 885 | 917 | 959 | 988 |
| PBIS | 709 | 795 | 882 | 914 | 957 | 987 |
| RAV | 624 | 779 | 882 | 914 | 960 | 989 |
| Validity | | | | | | |
| O-1 | 851 | 926 | 946 | 952 | 961 | 930 |
| TW | 789 | 896 | 905 | 931 | 913 | 888 |
| BIS | 829 | 911 | 948 | 952 | 958 | 946 |
| PBIS | 843 | 916 | 949 | 954 | 963 | 954 |
| RAV | 805 | 892 | 940 | 948 | 950 | 934 |

Note.--Decimal points have been omitted

studies, the increase in internal consistency reliability due to DOWS is not unconditional. Conventional scoring yields more reliable scores than DOWS procedures when the inter-item correlation is .16 or less. However, for more homogeneous tests ($r_{tet} \geq .25$) DOWS procedures do yield more reliable scores. TW appears to be the least effective of the experimental scoring procedures in increasing reliability. BIS is slightly superior to PBIS and RAV.

In general, DOWS procedures do not increase the validity of the scores. TW and RAV yield consistently less valid scores than conventional scoring (with one exception). BIS also yielded less valid scores except for tests with $r_{tet}$ of .25 and .81. On the other hand, PBIS yielded consistently more valid scores than conventional scoring for tests with $r_{tet} \geq .25$.

## CONCLUSIONS

Although the present study is limited in generalizability as to the effect of DOWS on validity and reliability, it appears clear from the results that comments regarding the effect of DOWS should not be made without taking into account inter-item correlation <u>and</u> the scoring procedures. While previous studies had concluded that DOWS yielded more reliable scores, this study has shown that increases in reliability are observed only for tests with certain characteristics, namely, inter-item correlations of .25 or higher. Similarly the effect of DOWS on validity seems to depend on inter-item correlation and the particular scoring method. That is, TW, BIS and RAV seem to have a negative effect on validity whereas PBIS appears to yield more valid scores than conventional scoring for tests with inter-item correlation of .25 or higher. The reasons why only PBIS, and not other scoring procedures, performed in this way are not clear. However, the finding that the effect was doubly-cross-validated (data for only one cross-validation group were reported in Table 1; see Bejar, 1973) should encourage further investigation.

From a practical standpoint however, the question remains whether DOWS is worth the effort. In order to answer that question, several factors must be taken into account. First, the hypothetical tests used were constructed in such a way that the alternatives formed a hierarchy of merit. Clearly this is the sort of test where DOWS would seem to perform optimally. Secondly, the increase in validity (for PBIS) was not very large, although consistent. What this means in practical terms is that in order to increase validity by a small amount the test constructor would have to construct items where the alternatives form a hierarchy, and then compute and cross-validate the option weights. Thirdly, the tests for which increases in reliability and validity were observed were those with inter-item correlations of .25 and higher. In practice, tests are rarely composed of items with such high inter-item correlations. Thus, it appears that the pay-off of DOWS under most circumstances would be limited indeed.

# REFERENCES

Bejar, I.I. An empirical investigation of the effect of differential weighting of alternatives on validity and reliability. Unpublished M.A. Thesis. University of Minnesota, 1973.

Davis, F.B. & Fifer, G. The effect on test reliability and validity on scoring aptitude and achievement tests with weights for every choice. Educational and Psychological Measurement, 1959, 19, 159-170.

Hendrickson, G.F. The effects of differential option weighting on multiple-choice objective tests. Report No. 93 Johns Hopkins University, 1970.

Hendrickson, G.F. & Green, B.F. Jr. Comparison of the factor structure of Guttman weighted vs rights-only-weighted tests. Paper presented at the meeting of the American Educational Research Association, Chicago, April, 1972.

Reilly, R.R. & Jackson, R. Effect of empirical option weighting on reliability and validity of the GRE. Research Bulletin 72-38, Princeton, N.J.: Educational Testing Service, 1972.

Sabers, D.L. & White, G.M. The effect of differential weighting on individual item responses on the predictive validity and reliability of an aptitude test. Journal of Educational Measurement, 1969, 6, 93-96.

Shoemaker, D.M. & Osburn, H.G. A simulation model for achievement testing. Educational and Psychological Measurement, 1970, 30, 267-272.

Wang, M.W. & Stanley, J.C. Differential weighting: A review of methods and empirical studies. Review of Educational Research, 1970, 40, 663-705.

# THE STRATIFIED ADAPTIVE COMPUTERIZED ABILITY TEST

David J. Weiss

Research Report 73-3

Psychometric Methods Program
Department of Psychology
University of Minnesota

September 1973

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

MKC
[illegible]
73-3

Security Classification

## DOCUMENT CONTROL DATA - R & D

*(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)*

| | |
|---|---|
| 1. ORIGINATING ACTIVITY *(Corporate author)*<br>University of Minnesota<br>Department of Psychology | 2a. REPORT SECURITY CLASSIFICATION<br>unclassified<br>2b. GROUP |

3. REPORT TITLE

The Stratified Adaptive Computerized Ability Test

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*
Technical Report

5. AUTHOR(S) *(First name, middle initial, last name)*

David J. Weiss

| | | |
|---|---|---|
| 6. REPORT DATE<br>September 1973 | 7a. TOTAL NO. OF PAGES<br>45 | 7b. NO. OF REFS<br>17 |

| | |
|---|---|
| 8a. CONTRACT OR GRANT NO.<br>N00014-67-A-0113-0029 | 9a. ORIGINATOR'S REPORT NUMBER(S)<br>Research Report 73-3<br>Psychometric Methods Program |
| b. PROJECT NO.<br>NR 150-343 | |
| c. | 9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)* |
| d. | |

10. DISTRIBUTION STATEMENT

Approved for public release; distribution unlimited.

| | |
|---|---|
| 11. SUPPLEMENTARY NOTES | 12. SPONSORING MILITARY ACTIVITY<br>Personnel and Training Research Programs, Office of Naval Research |

13. ABSTRACT

The stratified adaptive (stradaptive) test is described as a strategy for tailoring an ability test to individual differences in testee ability. Stradaptive test administration is controlled by a time-shared computer system. The rationale of the method is described as it derives from Binet's strategy of ability test administration and findings concerning peaked tests from modern test theory. The essential elements of stradaptive testing which are considered include the differential entry point, branching rules, and individualized termination criteria. Different methods of scoring the stradaptive test are discussed, as are the implications of individual differences in consistency of test responses within the stradaptive test record. A number of examples of the results of live stradaptive testing are presented and discussed. Implications of additional data derived from stradaptive test response records are considered and related to other psychometric concepts.

Security Classification

| 14. KEY WORDS | LINK A | | LINK B | | LINK C | |
|---|---|---|---|---|---|---|
| | ROLE | WT | ROLE | WT | ROLE | WT |
| testing | | | | | | |
| ability testing | | | | | | |
| computerized testing | | | | | | |
| adaptive testing | | | | | | |
| sequential testing | | | | | | |
| branched testing | | | | | | |
| individualized testing | | | | | | |
| tailored testing | | | | | | |
| programmed testing | | | | | | |
| response-contingent testing | | | | | | |
| automated testing | | | | | | |
| stradaptive test | | | | | | |

## Contents

# THE STRATIFIED ADAPTIVE COMPUTERIZED ABILITY TEST

Since the development of the first group ability test over a half-century ago, paper and pencil tests have dominated ability testing. Paper and pencil tests, which represent one strategy of measuring human abilities, consist of a limited number of test items organized in a specified manner which are presented to all testees in the same way. Testees proceed through the test items in approximately the order in which they are printed in the test booklet. The paper and pencil test is thus a highly standardized testing strategy which was developed to permit one administrator to test large numbers of testees simultaneously. However, the group paper and pencil test has a number of deficiencies (Weiss & Betz, 1973) which make it desirable to investigate other strategies of administering ability tests.

The availability of time-shared computer systems now makes it possible to implement a variety of new strategies for measuring abilities. Interactive computer systems, in which the testee can be presented with test items by the computer and respond to them on a typewriter keyboard, or by means of a light-pen, permit the psychometrician to develop ways of adapting, or tailoring, test items to each individual's estimated ability level. This is accomplished as a result of the computer's capacity to receive the testee's response to a test item, evaluate that response, consult a pre-determined set of rules to determine the next item to be administered, and to administer the chosen next item. In a time-shared computer system, one computer can administer such adaptive ability tests essentially simultaneously to a large number of testees.

In adaptive testing it is the "pre-determined set of rules" governing the choice of the next test item to be administered that differentiate the various strategies of computerized ability testing. In paper and pencil testing each item is administered in succession whether a testee answers an item correctly or incorrectly. In adaptive testing, choice of the next item to be administered is contingent upon whether the testee's response to a previous item, or a set of previous items, was correct or incorrect. A number of different strategies, or decision rules for choice of subsequent test items, have been proposed to implement adaptive testing (Weiss & Betz, 1973). Among these are two-stage, pyramidal, flexilevel, Bayesian and maximum likelihood approaches for tailoring or adapting a test to individual differences among testees.

While each of these available adaptive testing strategies has its advantages and unique characteristics (Weiss, 1973), logical considerations suggest that additional ways of moving a testee through an item pool might be desirable. This paper proposes one such new method, describes its rationale, and presents some examples based on actual computerized testing.

## "Peaked" Ability Tests

A peaked ability test is one in which all test items are very similar in difficulty. In the extreme case of peakedness, an ability test would have all items of the same level of difficulty. Thus, item difficulty would have no variance. Since this ideal condition is rather difficult to achieve in practice, operational peaked ability tests tend to have very low variances of their item difficulties, reflecting a set of test items distributed over a very narrow range of difficulty. The smaller the item difficulty variance, the greater the peakedness. When the range of the distribution of item difficulties in a test approaches the range of ability measured by that test, and there are an equal number of items at each level of difficulty, the distribution of item difficulties is said to be rectangular. Most commercial ability tests have distributions of item difficulties which lie between the extremes of the completely peaked test and the rectangularly distributed ability test. These tests tend to have item distributions which are approximately normally distributed across the ability continuum.

In a series of theoretical papers comparing completely peaked ability tests (i.e., tests composed of items of equal difficulty) with tests "administered" under a variety of adaptive testing strategies, Lord (1970; 1971a,b,c) reached one consistent conclusion: in terms of the precision of measurement, or the capability of responses to a set of test items to reproduce accurately the "true ability" of hypothetical testees, the 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 test was peaked. As the testee's ability deviated from the point at which the test was peaked, the measurement efficiency (i.e., the number of test items required to achieve a given degree of precision) of the peaked test diminished more rapidly than that of the adaptive tests. Figure 1 illustrates Lord's general finding in this series of studies. As Figure 1 shows, at some point on the ability continuum, usually plus or minus .50 to 1.0 standard deviations, the efficiency of the adaptive test becomes higher than that of the peaked test.

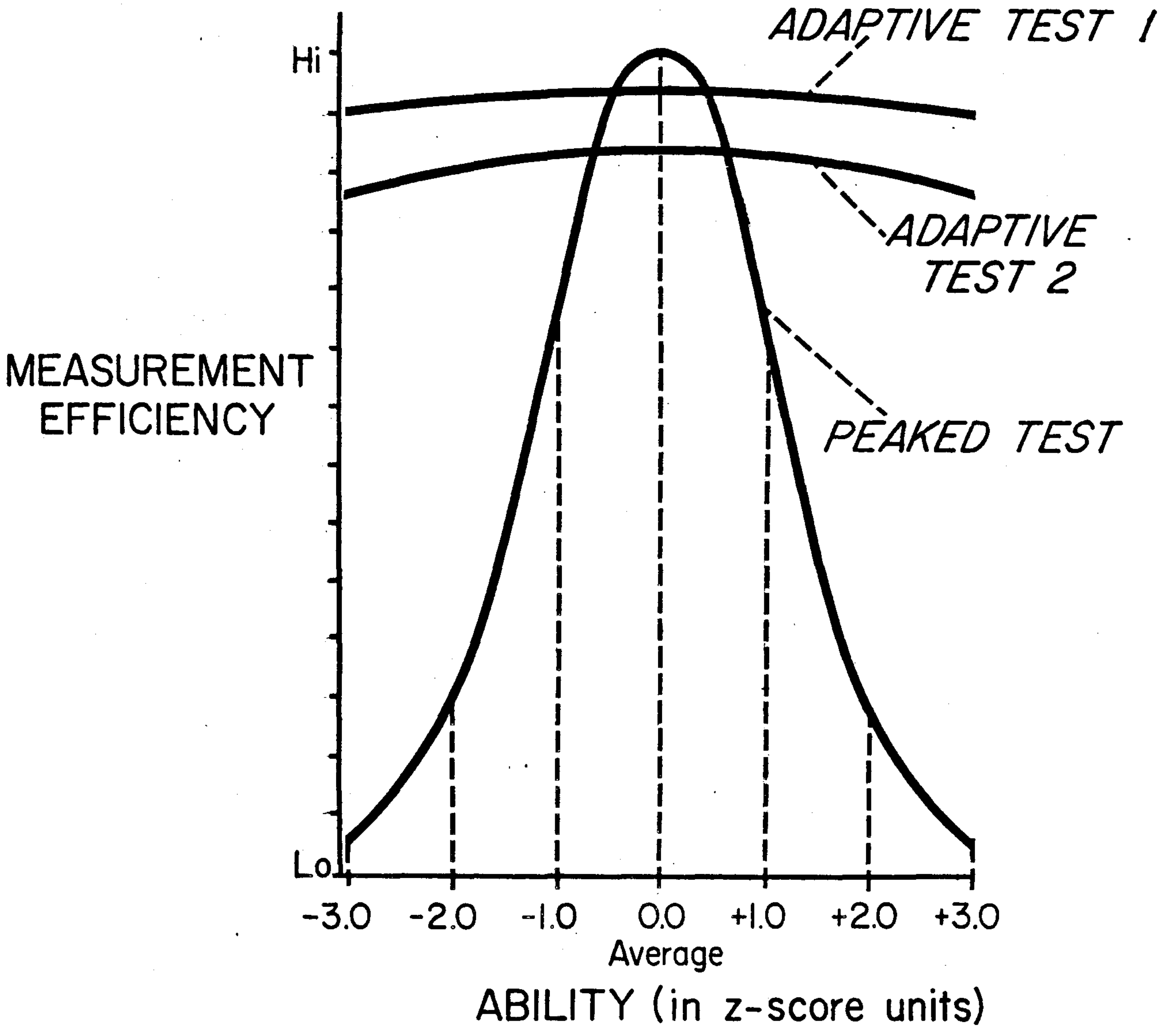


Figure 1. Efficiency of measurement as a function of ability level (after Lord, 1970; 1971a,b,c)

With increasing distance from the peaked point, the adaptive tests become more and more efficient in comparison to the peaked test. However, Lord's theoretical results did show that peaked tests can provide greater measurement efficiency than all adaptive tests studied thus far for up to about 70% of a population normally distributed around the peaked point of the test.

While Lord's theoretical analyses reflect an ideal set of conditions (i.e., all test items are of equal difficulty and equal discrimination), they are important enough not to be easily dismissed. Interpreted in another way, Lord's findings indicate that peaked tests provide most accurate measurement when the ability of the individual being measured is exactly equal to the difficulty level at which the test is peaked. His analysis is supplemented by the findings of information theory (e.g., Hick, 1951) which indicate that test items provide most information when the probability of a correct answer to a given test item is .50 for any individual. Thus, a test comprised of all items of .50 difficulty _for an individual_ would provide the most information about that individual's true ability level, and in Lord's terms, the most precise test score for him.

The important aspect of these findings from both test theory and information theory is that the test must be peaked at the individual's ability level for measurement to be most accurate. But ability level is not known in advance; it is the test's function to measure ability level. The typical solution to this problem is to peak tests at the estimated ability level of some _group_ of testees. Thus, a test designed to measure the abilities of college freshmen is peaked at the average ability level for college freshmen. Since testees always vary in ability, however, the precision of measurement of any individual's ability estimate derived from a peaked test will depend on the distance of his ability from the estimated mean ability of the group, as shown in Figure 1. Thus, the individual whose ability is at the group mean will have a test score of maximum precision. But individuals whose ability deviates from that mean will obtain ability estimates which are less precise, with precision decreasing with increasing distance from the mean. For individuals below the estimated mean ability level of the group, the test items will be too difficult. For these testees the probability of correctly answering the items will be less than .50; the items thus will provide less information on their true ability level. For individuals above the estimated mean ability level, the items will be too easy. Thus, their probability of a correct response will be greater than .50 and again, the test items will provide less information about the ability levels of those testees.

Following the administration of a peaked test, it is possible to tell if the test was appropriate for any given individual. If the test is peaked with items of average difficulty for a group of subjects, the difficulties of the items will be p = .50, i.e., half the group will have answered each item correctly. The appropriateness of that peaked test for any individual can be determined by the proportion of total items taken that he/she has answered correctly. A peaked test can be thought of as being most appropriate for an individual if he gets about half the items correct. Under these circumstances each item provides maximum information on that testee and his score has maximum precision. If an individual answers none of the items on a test correctly (or, if guessing is possible, operates at a chance level) or answers most or all the items in the test correctly, the test was inappropriate for that individual (Lord, 1971c). However, under conventional ability test administration procedures (i.e., paper and pencil tests), the appropriateness or inappropriateness of a test for any given individual can not be determined until after the test has been administered. For many uses of test information, such <u>post hoc</u> determination of appropriateness is too late; the obtained ability estimates may have associated with them very large errors which seriously reduce their utility in practical situations and frequently result in invalid uses of such test scores for practical decisions.

## Binet's Testing Strategy

Recognition that a single peaked test may not be appropriate for a given testee seems to have been implicit in Binet's early work in individual testing. That work resulted in the Stanford-Binet Scales (Terman and Merrill, 1960), which are still acknowledged by many as the "standard" of ability measurement. Binet's approach to ability measurement, rather than depending on a single test peaked at the average ability level of the children whose ability it was measuring, used a series of tests organized around the concept of "mental age." Test items at each of the "mental age" levels were peaked around a given mental age, and there was little overlap between mental ages. Items were included in a peaked "mental age" test if about 50% of the norm group of that chronological age gave correct answers to those items. In other words, the items in the test labelled "mental age 8.0", for example, would be those items answered correctly by approximately 50% of those aged exactly 8.0 years who were part of the norm group. A similar rationale was used to construct the tests peaked at each other "mental age" comprising the Binet test. The Stanford-Binet can thus be characterized not as one test but as a series of tests, each peaked at a given mental age and providing most accurate measurement for individuals at that mental age.

Binet's test administration procedure implicitly recognizes that peaked tests which do not permit the testee to obtain about half correct and half incorrect answers provide little information about his ability and therefore should not be administered to him. In administering the Stanford-Binet, the administrator estimates an "entry point" into the hierarchy of mental age peaked tests. The usual entry point consists of that mental age closest to the testee's chronological age; thus, the testee whose chronological age is 8 years, 1 month, will likely start with the test peaked at the 8.0 year level. The administrator is allowed flexibility, however. If it is hypothesized on the basis of prior information that the child is "bright" for his age, the 8 year 1 month child might be started at the 9.0 mental age test; conversely, the child who is expected to be "less bright" might be started at the test peaked at age 6.5.

Following determination of the "entry point" on the scaled peaked tests, the administrator administers the items of the entry-point peaked test and then moves to tests of lesser difficulty. Items are scored as test administration proceeds, with the administrator searching first for the testee's "basal age" and then for his "ceiling age." Binet's basal age is the peaked test at which the individual answers all test items correctly. These data provide no information on an individual's ability except that it is likely not to be lower than that mental age. Thus, it is assumed that if the testee were administered items from tests peaked at mental ages below the obtained basal age, he would provide correct answers to all of those items. If this assumption is correct, those items also will provide no information on the testee's ability level (they would all be too easy), thus nothing would be gained by administering them. The "basal age" therefore defines a "floor" below which further ability testing is unfruitful.

Similarly, the "ceiling age" provides an upper limit beyond which further testing is unnecessary and, in terms of testee motivation (e.g., frustration), might even reduce the accuracy of the test score. The "ceiling age" identifies the peaked test at which the testee obtains all incorrect answers. Like the basal age test, in terms of information theory the test responses provide no information. The ceiling age simply indicates that the individual's ability is somewhere below that level, but it does not indicate where on the ability continuum the individual is likely to be located. It is also assumed that all peaked tests above the ceiling age will likely produce the same results as the ceiling age test, i.e., all responses would be incorrect, and therefore the tests would provide no information on the testee's ability level.

Once the administrator has determined a testee's basal age, testing proceeds through tests of higher difficulty until the ceiling age is identified. It is the peaked tests within the limits defined by the basal and ceiling ages that will likely provide meaningful information on a testee's ability level. The totality of test items between any testee's basal and ceiling ages will provide accurate measurement for that individual; for another testee with different basal and/or ceiling levels a different set of test items will provide maximum information on his ability level. If the test is properly unidimensional for a given individual, and administration conditions are optimal, the proportion correct at each mental age level from the basal age through the ceiling age should show a regular decrease. If there were a very large number of mental age peaked tests between the basal and ceiling ages, proportion correct on these tests would vary from 1.00 at the basal age, through a test on which the individual answers approximately .50 of the items correctly, to .00 correct at the ceiling age. It will be noted that the area between the basal and ceiling ages includes a peaked test (at least theoretically) of maximum measurement efficiency, i.e., a peaked test on which the individual answers 50% of the items correctly.

Assuming that the item pool is relevant for each individual (i.e., they are from the culture on which the test was normed) and that it is unidimensional for each testee, the Stanford-Binet is the only test which has this characteristic--measurement of any individual's ability is confined to that area of the ability continuum which provides, over all test items administered, maximum average information per test item. The Stanford-Binet should, therefore, provide scores of more nearly constant precision of measurement than tests which do not have this adaptive feature--the capability of "searching out" the individual's ability level among a series of scaled peaked tests. Perhaps it is this characteristic of the Binet tests which has made them the standard of comparison for other ability tests.

Thus, by adapting selection and administration of peaked tests to the individual being measured, Binet's concept of ability testing seems to anticipate Lord's later theoretical findings concerning the efficiency of peaked tests. The individual administration of the Binet tests, however, introduces other sources of score variance which attribute error to the measurements obtained (Weiss & Betz, 1973). In addition to the unreliability due to scoring, administrator effects such as sex and race and other characteristics of the administrator and surrounding conditions serve to offset the increases in precision of measurement gained from the adaptive strategy of test administration.

With the current availability of time-shared computers for use as test administration devices, it is now possible to minimize the effects of the administrator variables which affect test scores, and at the same time utilize Binet's insights, with some improvements, in the ability measurement process. The stratified adaptive (STRADAPTIVE) computerized test is proposed as a means of obtaining ability test scores with nearly constant precision across a wide-ranging group of testees, building on the logic of Binet's test administration procedure and implementing Lord's theoretical findings and those available from information theory.[1]

## The STRADAPTIVE Test

The stradaptive test, like Binet's testing strategy, operates from a pool of items stratified by difficulty level, or organized into a set of scaled peaked tests. Each testee begins at a difficulty level estimated to correspond to his ability level, also following Binet's strategy. By using any of a number of branching procedures, the stradaptive test moves the testee through items of varying levels of difficulty in search of a region of the item pool which will provide maximum information about his ability level. The branching process leads to the identification of a "basal stratum" and a "ceiling stratum". Testing can be terminated when the ceiling stratum is reached. Each of these characteristics of stradaptive testing is considered below in detail.

[1] The term "stradaptive" is used rather than "stratified" to differentiate this approach from Cronbach's (Cronbach, Gleser, Nanda & Rajaratnam, 1972) conception of stratified tests, which are based on the idea of sampling test items from a stratified universe in which test items are classified by content, task, or difficulty.

## Item Pool Structure

The stradaptive test requires an item pool stratified by the difficulty levels of the constituent test items. A stratified item pool is one in which items are organized into a series of tests peaked at different difficulty levels. The pool should be known or assumed to be unidimensional. It will be shown below, however, that unidimensionality of the pool might not be evident for some testees; but the pool should be unidimensional for most testees in order to provide the most constant precision of measurement. The steps in developing an item pool for a stradaptive test include the following:

1. Administer a large number of items measuring the same ability to a large group of subjects. The subjects should be representative of the wide-ranging population for which the stradaptive test is intended. The size of the original item pool will depend on the quality of the items used and the target size of the final stratified item pool. While the optimal size of the stradaptive item pool is yet to be determined, adequate results have been obtained with about 200 items in the final pool. Likewise, no information is as yet available on the required number of subjects in the norming item pool. Naturally, a larger norming group will result in more stable item parameter estimates.

2. Derive item discrimination and item difficulty estimates for the items administered to the norming group. These parameters can be either traditional item parameters (proportion correct, item-total score correlations) or parameters derived from modern test theory using normal ogive item assumptions or logistic item functions (Lord & Novick, 1968). Items with very low discriminations should be eliminated.

3. Organize the item pool into a number of independent strata by difficulty level, where each stratum is in effect, a peaked test of some number of items. There should be no overlap in item difficulties between the strata. The number of strata developed from an item pool, or the number of peaked tests available, depends on the size of the original item pool. The larger the number of strata the more likely the obtained ability tests will have equal precision across a group of testees of wide-ranging ability, since the peaked tests

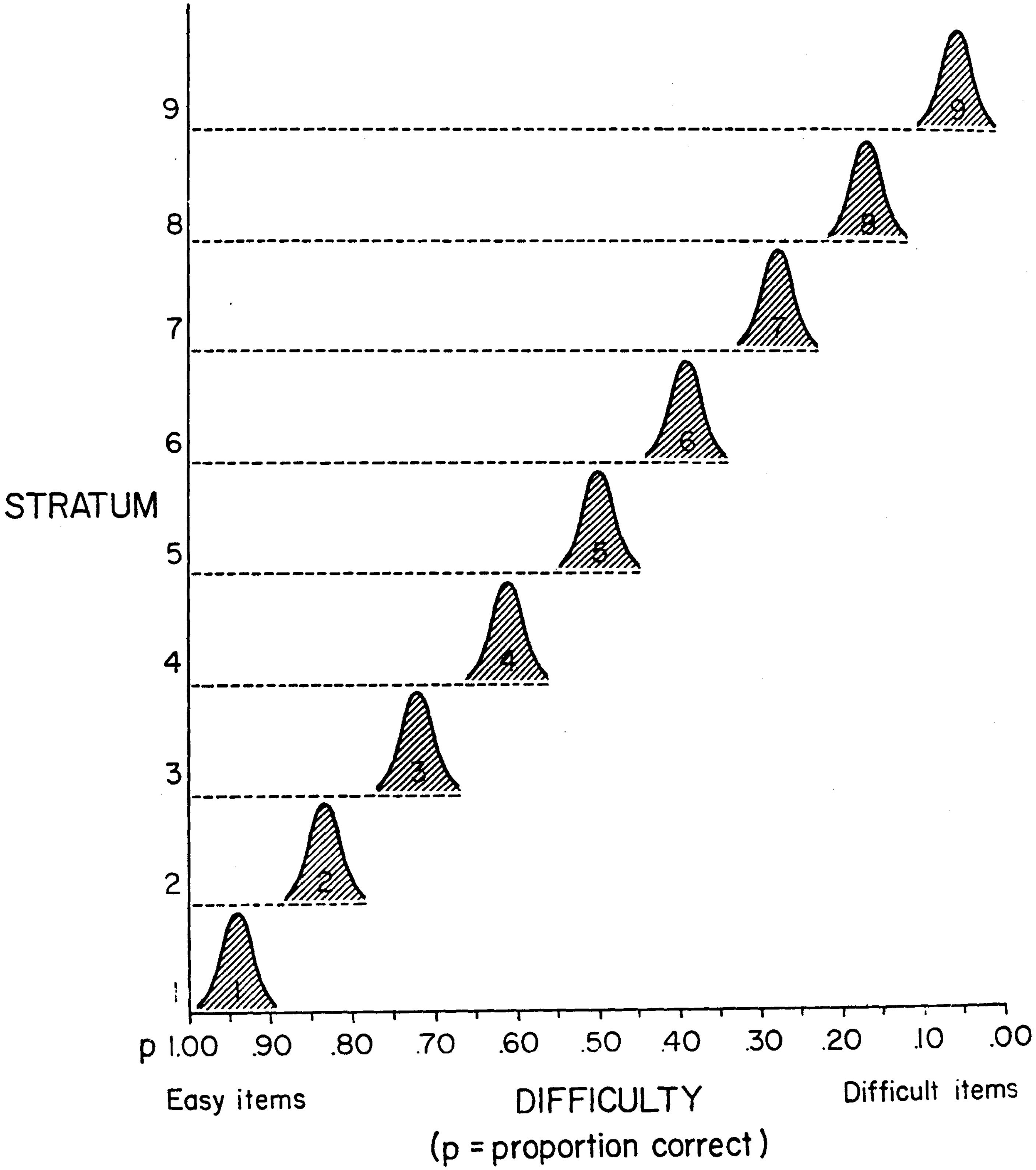


Figure 2. Distribution of items, by difficulty level, in a Stradaptive Test

are more likely to exactly match each testee's ability level. A minimum of nine or ten strata seems to be appropriate, since that number of strata seems to provide a good range of coverage of abilities without requiring very large item pools. The question is, of course, open for considerable further investigation.

The number of items at each stratum will vary with both the size of the original item pool and with the number of strata to be developed. A minimum of ten to fifteen items at any given stratum appears to be appropriate. There need not be an equal number of items at the various strata; experience suggests that the middle and lower difficulty strata might require more items than those at the upper extremes.

4. The items within each stratum should be arranged in decreasing order of item discrimination, if item discrimination indices were derived from analyses on the total norming group, as differentiated from indices computed on sub-groups based on ability levels. Since at the earlier stages of testing (i.e., the first few items at each stratum) items must discriminate across a wider range of abilities, item discriminations based on a group of wide-ranging ability will be more appropriate. On the other hand, at the later stages of testing when testing is confined to only a narrow range of abilities (i.e., within 2 or 3 of the available strata), items need not be able to discriminate on a group of wide-range ability. Rather, item discriminations should be based on discrimination indices derived from closely contiguous levels of ability. Thus, items with relatively low discrimination indices on the total group might be capable of discriminating between contiguous strata at the later stages of testing (Paterson, 1962; Bryson, 1971).

The result of this process of structuring the item pool is shown diagrammatically in Figure 2. The hypothetical stradaptive item pool shown in Figure 2 contains nine strata. Each stratum consists of a subset of items peaked around a different difficulty level, with the difficulty level increasing with each successive stratum. Thus, stratum 1 consists of a sub-set of very easy items distributed approximately normally around a difficulty level of p = .94, with items varying in difficulty from p = .99 to p = .89; stratum 1, therefore, represents a very easy peaked test. Stratum 2 consists of a set of items peaked at a difficulty level slightly higher than those of stratum 1; stratum 2 items are peaked at about p = .83 and vary from p = .88 to p = .78. Stratum 9 is a difficult test with items varying in difficulty from p = .01 to p = .11 and peaked at p = .06. Note that the item distributions in Figure 2 do not overlap between strata.

Table 1 shows an operational stradaptive item pool. The pool consists of 229 items grouped into 9 difficulty strata. The number of items at each stratum varies from 10 at stratum 9 (the most difficult peaked test) to 36 at strata 2 and 3. Items were selected from a larger pool of about 500 items on which normal ogive transformations of item discriminations (a) and difficulties (b) had been previously computed using estimates of Lord's (Lord & Novick, 1968) normal ogive item parameters. To construct the item pool, the range of item difficulties from +3.00 standard deviations to -3.00 standard deviations was divided into 9 equal parts. All items from the larger pool were included in the stradaptive item pool if their normal ogive discrimination parameters were a = .30 or above (with the exception of the tenth item at stratum 9 which was included to increase the number of items at that stratum to 10).[2]

The 9 strata in Table 1 are essentially nine peaked tests varying in average difficulty from -2.65 to +2.62. The most difficult peaked test (stratum 9) is composed of 10 items peaked at b = 2.62, varying from the most difficult item at b = 3.11 to the easiest item in that stratum at b = 2.32. Stratum 8 is a slightly less difficult peaked test with average b = 2.01 and with the 15 items varying in difficulties from b = 2.31 to b = 1.65. Within each stratum items are ordered by discrimination; for stratum 9 the first item has a discrimination of a = .84, and the last item at that stratum has a discrimination of a = .21. Similar patterns are obvious for the other strata. The greater number of items at the middle and lower level of difficulties reflects the composition of the original item pool from which these items were selected. However, in actual testing with the stradaptive test it has become evident that successful testing for many subjects requires the availability of a larger pool of items at the middle and lower ranges of difficulty.

## Operationalizing the Stradaptive Test

Entry point. The stradaptive test permits the use of differential entry points for beginning testing for different individuals. While it is not necessary to use

[2] A further exception is item 19 at stratum 4, which has a discrimination of .27; that item was included in the pool by error.

## Table 1

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

| | (easy) | | | | | | | Stratum | | | | | | | | | (difficult) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | | 2 | | 3 | | 4 | | 5 | | 6 | | 7 | | 8 | | 9 | |
| Item Difficulties | | | | | | | | | | | | | | | | | | |
| Hi | -2.39 | | -1.64 | | -1.01 | | -0.34 | | 0.33 | | .98 | | 1.63 | | 2.31 | | 3.11 | |
| Lo | -2.98 | | -2.32 | | -1.63 | | -1.00 | | -0.28 | | .34 | | 1.00 | | 1.65 | | 2.32 | |
| Mean | -2.65 | | -1.92 | | -1.29 | | -0.63 | | .02 | | .65 | | 1.33 | | 2.01 | | 2.62 | |
| No. of items | 35 | | 36 | | 36 | | 30 | | 25 | | 19 | | 23 | | 15 | | 10 | |
| Item Number Within Stratum | b | a | b | a | b | a | b | a | b | a | b | a | b | a | b | a | b | a |
| 1 | -2.42 | 3.00* | -1.99 | 1.76 | -1.51 | 1.40 | -.70 | 1.82 | -.05 | 1.31 | .73 | .98 | 1.07 | .72 | 1.89 | .85 | 2.95 | .84 |
| 2 | -2.42 | 3.00 | -1.78 | 1.54 | -1.23 | 1.35 | -.73 | .92 | .14 | 1.07 | .34 | .91 | 1.49 | .62 | 2.03 | .64 | 2.47 | .48 |
| 3 | -2.45 | 3.00 | -2.22 | 1.52 | -1.08 | 1.23 | -.52 | .86 | -.13 | .98 | .65 | .77 | 1.33 | .60 | 1.93 | .57 | 2.61 | .43 |
| 4 | -2.45 | 3.00 | -1.68 | 1.46 | -1.33 | 1.16 | -.68 | .86 | .15 | .97 | .79 | .70 | 1.54 | .58 | 2.31 | .54 | 2.86 | .42 |
| 5 | -2.72 | 3.00 | -1.87 | 1.43 | -1.34 | 1.02 | -.59 | .83 | -.08 | .91 | .79 | .63 | 1.11 | .56 | 1.79 | .50 | 2.35 | .42 |
| 6 | -2.72 | 3.00 | -1.92 | 1.23 | -1.10 | .99 | -.75 | .82 | .16 | .86 | .49 | .56 | 1.40 | .55 | 2.04 | .49 | 2.67 | .42 |
| 7 | -2.72 | 3.00 | -1.88 | 1.14 | -1.42 | .92 | -.57 | .77 | -.21 | .86 | .42 | .55 | 1.17 | .52 | 1.79 | .49 | 2.32 | .38 |
| 8 | -2.66 | 1.79 | -2.13 | 1.10 | -1.21 | .91 | -.85 | .75 | -.25 | .86 | .98 | .52 | 1.30 | .52 | 1.88 | .45 | 2.37 | .34 |
| 9 | -2.54 | 1.59 | -1.64 | 1.08 | -1.06 | .89 | -.47 | .71 | .21 | .86 | .37 | .50 | 1.38 | .51 | 2.07 | .43 | 3.11 | .32 |
| 10 | -2.81 | 1.48 | -2.22 | 1.07 | -1.34 | .89 | -.40 | .68 | .16 | .83 | .46 | .49 | 1.44 | .49 | 2.13 | .42 | 2.50 | .21 |
| 11 | -2.46 | 1.29 | -1.67 | 1.02 | -1.31 | .87 | -.90 | .67 | -.23 | .81 | .46 | .48 | 1.31 | .44 | 2.31 | .40 | | |
| 12 | -2.78 | 1.26 | -1.71 | .99 | -1.10 | .77 | -1.00 | .67 | .30 | .78 | .65 | .48 | 1.25 | .43 | 1.65 | .39 | | |
| 13 | -2.47 | 1.16 | -2.26 | .98 | -1.55 | .77 | -.69 | .66 | .08 | .76 | .78 | .45 | 1.00 | .42 | 1.82 | .36 | | |
| 14 | -2.43 | 1.01 | -2.21 | .96 | -1.07 | .76 | -.81 | .66 | -.28 | .75 | .71 | .44 | 1.00 | .40 | 2.26 | .35 | | |
| 15 | -2.86 | 1.01 | -1.66 | .93 | -1.43 | .76 | -.56 | .66 | .24 | .66 | .65 | .43 | 1.26 | .39 | 2.18 | .34 | | |
| 16 | -2.94 | .96 | -1.65 | .92 | -1.40 | .75 | -.58 | .66 | .33 | .53 | .62 | .41 | 1.36 | .37 | | | | |
| 17 | -2.83 | .94 | -1.65 | .82 | -1.15 | .73 | -.84 | .65 | -.23 | .43 | .83 | .37 | 1.24 | .36 | | | | |
| 18 | -2.74 | .93 | -2.32 | .80 | -1.42 | .71 | -.85 | .64 | .09 | .43 | .75 | .37 | 1.60 | .35 | | | | |
| 19 | -2.89 | .91 | -1.80 | .77 | -1.63 | .71 | -.41 | .27 | .15 | .42 | .92 | .37 | 1.21 | .35 | | | | |
| 20 | -2.54 | .88 | -1.80 | .76 | -1.47 | .67 | -.94 | .60 | -.09 | .41 | | | 1.47 | .34 | | | | |
| 21 | -2.55 | .79 | -1.93 | .74 | -1.60 | .66 | -.41 | .59 | -.26 | .40 | | | 1.61 | .34 | | | | |
| 22 | -2.81 | .74 | -2.28 | .70 | -1.33 | .62 | -.89 | .53 | .08 | .39 | | | 1.63 | .32 | | | | |
| 23 | -2.50 | .68 | -1.83 | .66 | -1.04 | .58 | -.52 | .48 | .09 | .37 | | | 1.36 | .31 | | | | |
| 24 | -2.82 | .67 | -1.74 | .63 | -1.17 | .57 | -.58 | .48 | -.04 | .35 | | | | | | | | |
| 25 | -2.54 | .67 | -1.70 | .59 | -1.27 | .56 | -.39 | .40 | .12 | .32 | | | | | | | | |
| 26 | -2.50 | .66 | -2.19 | .56 | -1.07 | .56 | -.36 | .40 | | | | | | | | | | |
| 27 | -2.51 | .64 | -1.89 | .52 | -1.02 | .54 | -.58 | .40 | | | | | | | | | | |
| 28 | -2.39 | .62 | -2.20 | .50 | -1.01 | .52 | -.38 | .38 | | | | | | | | | | |
| 29 | -2.58 | .57 | -1.71 | .47 | -1.31 | .52 | -.34 | .32 | | | | | | | | | | |
| 30 | -2.98 | .56 | -2.21 | .44 | -1.30 | .52 | -.67 | .30 | | | | | | | | | | |
| 31 | -2.73 | .52 | -2.08 | .42 | -1.19 | .52 | | | | | | | | | | | | |
| 32 | -2.77 | .50 | -1.80 | .42 | -1.57 | .49 | | | | | | | | | | | | |
| 33 | -2.68 | .48 | -1.82 | .42 | -1.26 | .44 | | | | | | | | | | | | |
| 34 | -2.56 | .44 | -2.12 | .41 | -1.59 | .38 | | | | | | | | | | | | |
| 35 | -2.95 | .41 | -1.92 | .32 | -1.35 | .34 | | | | | | | | | | | | |
| 36 | | | -1.84 | .31 | -1.08 | .32 | | | | | | | | | | | | |

*Discriminations (a) were arbitrarily set to 3.00 when the biserial item-test correlation was .90 or higher.

differential entries, i.e., all testees can begin with the same test item, the differential entry point has at least two major advantages. First, beginning testing at different strata for different individuals might save time in testing in terms of the number of items administered to a given individual. Thus, if it is known or suspected that a given testee is likely to be high on the ability to be measured, say 1.5 standard deviations above the mean, it would be wasteful of the testee's time to begin testing with an item of average difficulty. Use of a differential entry point for this individual might save time by eliminating the administration of three or four unnecessary items. The time saving would increase as the individual's estimated ability deviated from an arbitrary fixed entry point.

The second major advantage of using a differential entry point for beginning testing involves the testee's motivation to continue testing or to do well. Beginning an individual of low ability at an item of median difficulty will almost insure that the first several items taken will be too difficult for him; a frustration or anxiety reaction might occur which could adversely affect his performance on the remainder of the test items. Conversely, administering items of median difficulty to an individual of high ability might cause a boredom or "irrelevance" reaction which could then affect his performance on the entire test.

It thus appears to be desirable to begin the stradaptive test at some point estimated to be approximately representative of the individual's ability level on the trait being measured. Two sources of entry point estimates are possible. First, the computer could have stored information on an individual which might be useful as entry point information. For example, if the stradaptive test is being used to measure verbal ability, such information as scores on other verbal ability tests, grades in English courses, grade point average, or simply number of years of formal schooling completed could be stored in the computer. Once the testee identifies himself to the computer by name or identification number, the computer would retrieve the appropriate information from his file and, based on known or estimated relationships between the prior information and test performance, determine the entry point on the ability continuum for that testee.

The testee himself is a second important source of entry point information. Rather than consulting actual records on the testee, it might be fruitful to ask testees for the information necessary to derive entry points.

Figure 3 shows two such entry point questions currently in use for stradaptive testing of verbal ability. The top half of Figure 3 is an entry point question for use with college students. In constructing the entry point estimate it was assumed that college grade point average (GPA) had a roughly positive and linear relationship with verbal ability. Individuals who answer in the first category, 3.76 to 4.00, enter the stradaptive test at stratum 9; individuals who indicate that their GPA's are between 2.51 and 2.75 enter the stradaptive test at stratum 4.

The bottom half of Figure 3 shows a different entry point question asked of the testee. This entry point information was developed for use with a group of inner-city high school students who could not be assumed to know their GPA and might also prove to be useful in a non-school testing situation. It is based on the assumption that the testee has a fairly good knowledge of his level of ability in comparison to his peers. Whether or not the testee can make a good estimate of his ability can be determined by the results of the stradaptive testing. The only effect of a poor estimate of a testee's entry point is that he will be administered a few more test items than would otherwise be necessary to measure his ability adequately. In any case, the stradaptive test is designed to converge upon the testee's level of ability regardless of the adequacy of the entry point. Thus, entry point information need only be very roughly related to the ability being measured.

Branching. The stradaptive test permits the use of virtually any branching rule for moving from an item at one stage to one at the next. Branching in the stradaptive test occurs between strata, therefore no pre-determined item branching network exists for the stradaptive test. The simplest branching rule is an "up-one/down-one" procedure. If a testee answers an item correctly, he is routed to an item at the next more difficult stratum; if he answers incorrectly he is routed to an item at the next easier stratum of difficulty. Other branching rules are also possible. For example, a correct response can lead to an item one stratum higher in difficulty, while an incorrect response can branch downward two strata. Such a rule might be adopted either where the opportunity for guessing may allow the testee to answer a number of items correctly solely by chance, or where it is desired to administer a very easy item (with a high probability of a correct answer for a given individual) following an incorrect response in order to prevent the testee from becoming discouraged.

Figure 3

## Stradaptive Test Entry Point Questions

### College Students

In which category is your cumulative GPA to date?

|  |  | Entry Stratum (not seen by student) |
|---|---|---|
| 1. | 3.76 to 4.00 | 9 |
| 2. | 3.51 to 3.75 | 8 |
| 3. | 3.26 to 3.50 | 7 |
| 4. | 3.01 to 3.25 | 6 |
| 5. | 2.76 to 3.00 | 5 |
| 6. | 2.51 to 2.75 | 4 |
| 7. | 2.26 to 2.50 | 3 |
| 8. | 2.01 to 2.25 | 2 |
| 9. | 2.00 or less | 1 |

Enter the category (1 through 9) and press the return key.

### Non-College Students

Everybody is better at some things than others.... Compared to other people, how good do you think your vocabulary is?

|  |  | Entry Stratum (not seen by testee) |
|---|---|---|
| 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 are better than (in vocabulary).

If it is desired to obtain a fairly quick estimate of the testee's "ceiling stratum" (i.e., the stratum at which he gets all items incorrect) the tester might use different branching rules at different stages of testing. At the earlier stages of testing, he might use an "up-two/down-two" rule in order to more quickly arrive at a narrower range of strata in which the testee's ability is likely to fall. Then, after perhaps the tenth stage of testing (i.e., ten items have been administered), the tester might adopt an "up-one/down-one" procedure which would concentrate item administration within the narrower range of strata (e.g., 2 or 3) estimated to include the testee's actual ability level.

The stradaptive test also allows for differential response option branching, as suggested by Bayroff (Bayroff, Thomas & Anderson, 1960). In this procedure, incorrect response alternatives in a multiple choice (or, for that matter, a free-response) test are graded in terms of the extent to which they show partial knowledge. A correct response always leads to the same upward branching decision. When an item is answered incorrectly, the step size of the downward branch (i.e., the number of strata branched over) is a function of the "incorrectness" of the chosen distractor. For example, a "very wrong" answer (e.g., a response given only by testees of very low ability) might lead to a downward branch of three steps; a response which is closer to being correct might result in branching two strata downward; while choice of the most plausible incorrect answer would branch the testee only one stratum down in difficulty. Such differential response option branching should permit more rapid identification of an individual's actual ability level, leading to a reduction in the time needed for the assessment of a particular ability.

For individuals whose abilities are at or near the highest or lowest stratum in the stradaptive item pool, there may be instances where items at higher or lower difficulty strata will not be available. In these cases, it will be necessary to administer successive items at the same stratum in place of the optimal items at higher or lower strata.

Termination. A unique feature of the stradaptive test is its individualized termination rule. In contrast to two-stage tests, all the pyramidal models, and the flexilevel test (see Weiss & Betz, 1973, for research on these strategies, and Weiss, 1973, for detailed descriptions of each), all of which administer a fixed and pre-determined number of items to each individual testee, the stradaptive test permits the number of items administered to each testee to vary. While both Owen's (1969, 1970) Bayesian adaptive testing strategy and Urry's (1970) maximum likelihood strategy do permit an individualized number of test items, both of these strategies require restrictive assumptions about the hypothesized shape of the underlying ability distribution, and necessitate sophisticated mathematical calculations which might be difficult or time-consuming to implement on some computer systems. The stradaptive test, while retaining the individualized number of items, makes no assumptions about the shape of the ability distribution and requires no complex calculations.

As indicated above, the stradaptive test can be conceived of as a search for the peaked tests most appropriate for an individual testee. These peaked tests, which provice maximum information on a testee's ability level, can be identified, after the fact, as tests on which the testee answered about 50% of the items correctly, if guessing is not a factor. A peaked test is inappropriate if the testee answers all items correctly or all items incorrectly. Thus, the objective of the stradaptive test is to locate the region of the item pool in which measurement efficiency will be maximum for any individual.

This objective can be realized by a simple accounting procedure. Regardless of the branching rules used, the computer simply keeps track of 1) the number of items administered at each stratum and 2) the number of items answered correctly at that stratum. After each item has been answered, the ratio of these two values, or the proportion correct at each stratum, is computed. Prior to administering the next item, the termination criterion is checked to determine whether it has been met. If the criterion has been met, testing is stopped and the individual's response record is scored. If not, an additional item is selected using the branching rules previously chosen for testing. That item is administered and scored, the proportion of items correct at each stratum is computed, and the termination criterion again checked. Testing continues until the termination criterion is met.

One logical criterion for terminating stradaptive testing involves identifying the lowest (i.e., easiest) stratum at which the individual is answering at a chance level. Thus, the stradaptive test can be viewed as a search for the testee's "maximum" level of performance on that set of test items. In a multiple choice test the chance level is determined by 1/c, where c is the number of response choices in each test item. Thus, for 5-alternative multiple choice items, answering 1 (or zero) out of

5 items correctly at a given stratum would indicate chance responding. Using such a termination rule, then, testing would continue until a stratum is identified at which the testee has responded at chance or below, provided that, say, five items have been administered at that stratum. The last condition is necessary to avoid the situation where a testee answers the first one or two items at a given stratum incorrectly, but would answer correctly well above chance levels if administered enough items at that stratum. Variations in the minimum number of items required at any stratum before the proportion correct is used to check the termination criterion will probably result in stradaptive test scores with varying degrees of precision and stability. For example, requiring a larger number of items will probably result in fewer inappropriately early terminations, while decisions made on smaller numbers of items within a stratum might result in some artifactually early terminations after which further testing may have led to higher ability scores.

Conceptually, then, the tester can control the degree of precision of the ability estimates derived from stradaptive testing by manipulating the termination criterion in one of two ways. First, he can require that a larger number of items be administered at the ceiling stratum before the termination criterion is evaluated for an individual. Secondly, the tester can directly manipulate the confidence level of the termination decision. This can be accomplished by directly positing an hypothesis of a proportion of correct responses of, say, $p = .20$. The obtained proportion of correct responses (for any specified number of items) at a given stratum can then be tested against the hypothesized value by standard hypothesis test-procedures. This would involve either a binomial expansion given $p$, $q$ and $N$ (the number of items administered), or the computation of a confidence interval around the obtained proportion of correct responses using the same parameters. The alpha value associated with the test of hypothesis, or the confidence level of the confidence interval, could be chosen in advance by the tester as a way of controlling the precision of the obtained ability estimate. Testing would then continue until the data at any stratum failed to reject the hypothesis of chance responding (e.g., $p = .20$), or until the computed confidence interval included the hypothesized chance value. As the number of test items at the termination stratum increased, the power of the statistical test would also increase, thereby likely increasing precision of measurement and such practical criteria as test-retest stability of the ability estimates.

The proposed termination rule is applicable to multiple choice test items with a constant number of response choices, to true false test items, and to free-response test items. For four-choice test items, the pseudo-chance level is .25, for seven-choice items it is 1/7 or .14, and for true-false items it is .50. For free-response items, the termination criterion becomes the lowest stratum at which the individual answers no items correctly. Thus, when guessing can be completely ruled out, the stradaptive test would continue as long as an individual gets any items correct at strata of increasing difficulty. This termination criterion is identical to Binet's "ceiling age."

Implementation of the "lowest chance stratum" termination rule yields interesting results in actual stradaptive testing with an "up-one/down-one" branching rule. In general, for the majority of individuals these procedures identify a "basal stratum", i.e., a stratum at which all items are answered correctly, and a "ceiling stratum", i.e., the least difficult stratum at which the testee responds at a chance level. In between these two limiting strata, the proportion correct on each stratum will vary between 1.00 and the chance level (.20 or less) and will decrease fairly systematically from the basal to the ceiling stratum. This pattern is evident even when a relatively small number of items has been administered. Specific examples will be given below.

For some individual testees, inconsistency in their response records will occasionally cause the stradaptive pool to exhaust the supply of test items at some stratum. Thus, for a variety of reasons (e.g., motivation, fatigue, inappropriateness of the item pool for that testee), some individuals will fail to reach a termination criterion at a given stratum before exhausting the item pool at that stratum. When this occurs, the branching procedure can be modified to eliminate downward branching but to continue upward branching. Thus, following a correct response the testee would be presented with an item at the next higher stratum, but following an incorrect response an item at the same stratum would be administered if the next lower stratum is exhausted. This procedure will lead to a very rapid identification of the testee's ceiling stratum, at the expense of the probable positively reinforcing value of alternating difficult and easier test items.

## Scoring

Since the stradaptive test adapts item presentation to characteristics of the individual being tested, the "number correct" score used almost universally for conventional tests is inappropriate. Number correct is inappropriate because the number of items administered to each individual will vary; some individuals reach termination in 11 or 12 items, while others require 30 or 40 items to safisfy the termination criterion. It might be expected, therefore, that determining the proportion of items correct for any testee would be an appropriate method of scoring the stradaptive test. Computing the proportion correct would account for individual differences in the number of items administered yet convey the same information as the number correct score.

However, this reasoning fails to take into account the fact that in the stradaptive test, item difficulties are tailored to the individual's ability level through the branching procedure. The end result of the branching procedure is to identify a subset of items on which the individual obtains about 50% correct responses. In the later stages of stradaptive testing, when the testing procedure begins to converge on an individual's ability level, each time an item is answered correctly the testee receives a more difficult item (at the next higher stratum). Because that item is likely to be too difficult for him, he will probably answer it incorrectly and will therefore receive an easier item. Since he is likely to get that item correct, the process will be repeated and the testee will approximately alternate between easier items and more difficult items until the termination criterion is reached. The proportion of items correct for an individual will, therefore, center around .50, with deviations from .50 due to inappropriate entry points, unusual testee-item pool interactions, guessing, or an item pool of inappropriate difficulty. Actual stradaptive testing results for over 300 testees show that the large majority of proportions correct vary from .40 to .60.

Since the number correct scores and their derivatives are inappropriate for stradaptive tests, new methods of scoring must be developed. Some methods that might prove satisfactory are suggested by the available research on pyramidal adaptive testing models (see Weiss & Betz, 1973, p. 20-35). Because of some similarities between the stradaptive models and the pyramidal tests (Weiss, 1973) some of these scoring methods can be applied to stradaptive testing. Other scoring methods are suggested by the logic of the stradaptive test itself, as it derives from Binet's approach to ability measurement. .

Following are a number of ways stradaptive tests can be scored. Most scoring methods assume that normal ogive difficulty parameters, or estimates thereof, have been computed for the items of the stradaptive test so that item difficulty data are on the same latent scale as ability estimates; in this way, item difficulties can be used to estimate the ability of persons correctly answering subsets of items. In using these parameters it is assumed that the items in the stradaptive item pool measure a single unidimensional continuum.

Highest item difficulty scores. These scoring methods are borrowed from the pyramidal testing models (e.g., Paterson, 1962; Bayroff & Seeley, 1967; Lord, 1970). They are all based on the "hurdle" conception of ability measurement; that is, the individual's ability level can be determined from the "height of the highest hurdle he can jump." The difficulty of an item is equivalent to the height of the hurdle; answering an item correctly implies jumping the hurdle. There are three variations of this score possible in the stradaptive test, with the third being unique to stradaptive testing:

1. Ability can be scored as the difficulty of the most difficult item answered correctly.

2. Since testing always terminates at an item at the ceiling stratum, ability can be measured as the difficulty of the "n+1th" item, or the item that would have been administered next if testing had not terminated. Thus, the individual who answers his final (nth) item correctly would obtain a higher ability estimate than the testee who answers the nth item incorrectly.

3. An individual's ability score can be conceived of as the difficulty of the most difficult item answered correctly below the testee's ceiling stratum.

A major weakness of these "highest item difficulty" scores is their probable unreliability, in terms of test-retest stability, if guessing is possible. Since in a multiple choice test it might be possible for a testee to obtain a correct answer above his true ability level solely by chance, the first two of these scoring methods would probably be unreliable. Method 2 would probably yield scores of somewhat lower reliability than method 1 since guessing would be more likely to occur on items at the testee's ceiling stratum. Method 3 is suggested as an alternative unique to the stradaptive test when guessing is expected to operate; since method 3 attempts to minimize the effects of chance successes, its results should be more stable than those of methods 1 or 2. When guessing is not possible, i.e., on free-response items, methods 1 and 3 will give similar results. Method 2 results will vary as a function of the adequacy of the termination rule.

Stratum scores. As indicated above, the stradaptive item pool can be considered to be a series of peaked tests graded in difficulty. Associated with each peaked test is a difficulty level, which can be characterized by the average difficulty of all items at a given stratum. That average diffculty level indicates the point on the underlying ability continuum at which each peaked test is peaked. It can, therefore, be used as an ability estimate for individuals in several ways, following the logic of scoring methods 1 through 3:

4. An individual's score is the difficulty level associated with the most difficult stratum at which he answered at least one item correctly.

5. The stradaptive test score can be determined from the difficulty level of the stratum of the n+1th item.

6. Test score is the difficulty level of the stratum just below the testee's ceiling stratum, i.e., the difficulty of the highest non-chance stratum reached.

These stratum scoring methods might result in somewhat more stable ability estimates than the "highest item" methods, since they would eliminate some of the variability due solely to variations in difficulties of specific items which would occur in methods 1 to 3. In using scoring methods 4 through 6, however, the number of possible scores will be equal only to the number of strata. Thus, when the number of strata is small, score variability will be severely decreased, leading to loss of information on individual differences and lowered correlations with other variables. The stratum scoring methods appear appropriate, therefore, only when the number of strata in the item pool is quite large (e.g., 25 or more).

Scoring method 6 also does not convey information on the proportion of items correct at the stratum just below the testee's ceiling stratum. At that highest non-chance stratum, one testee might answer 80% of the items correctly, while another might answer only 25% of the items correctly; using scoring method 6, both of these testees would obtain the same score even though their ability levels are probably different. It seems appropriate, therefore, to define an additional method of scoring, the "interpolated stratum difficulty score", which is designed to take account of the proportion correct data on individual testees at the highest non-chance stratum.

7. The interpolated stratum difficulty score can be defined as:

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

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

$p_{c-1}$ is the testee's proportion correct at the c-1th stratum.

and S is $\overline{D}_{c} - \overline{D}_{c-1}$, if $p_{c-1}$ is greater than .50,
or $\overline{D}_{c-1} - \overline{D}_{c-2}$ if $p_{c-1}$ is less than .50,

where $\overline{D}$ is the average difficulty of the designated stratum.

The interpolated stratum score assumes that the testee's ability lies at the mean of the difficulties of a peaked test (i.e., a stratum) if he answers exactly 50% of the items on that test correctly. If he answers very few of the items correctly, for example 25%, his ability is below the mean of that peaked test, tending toward the mean of the items at the next lower stratum. If the testee answers 80% of the items at a stratum correctly, his ability is above the mean of the peaked test and close to the lower range of ability measured by the items at the next most difficult stratum. Essentially, then, this scoring method interpolates the testee's ability level as a function of the distance between the relevant mean difficulties of the strata and the proportion of items answered correctly. In implementing the computations, if the cth or c-2th strata do not exist (i.e., are above or below the difficulties available in the item pool) the average difficulty of those hypothetical strata can be determined by adding or subtracting the constant or increment in difficulty between strata to the last actual average stratum difficulty available.

The interpolated stratum difficulty score, in addition to having the desirable characteristic of taking account of more of the information available from stradaptive testing, has the added advantage of increasing the range of scores possible over that available from the other stratum scoring methods.

Average difficulty scores. In an effort to compromise the probable unreliability of scoring methods 1-3 and the restricted range of methods 4-6, a number of average difficulty scores appear to be logically sound:

8. An individual's score can be determined as the average difficulty of all items answered correctly.

This method continues the "hurdle" analogy of ability scoring, but attempts to balance out chance factors by using an average. A major deficiency of this scoring method is that scores will be affected by inappropriate entry points. If the entry point is too low the testee will be presented with, and probably answer correctly, a number of items below his true ability level. His ability estimate will, therefore, be lower than it should be. An inappropriately high entry point will result in the administration of a number of items which are too difficult for a given testee. The administration of these difficult items might increase the probability of chance successes and thereby artifactually raise test scores based on this method of scoring.

9. Ability can be scored as the average difficulty of all items correct between (but not including) the basal stratum (100% correct) and the ceiling stratum (chance responding).

Thus, the "routing items", those items resulting from too high or too low an entry point, will not be scored in this method. Therefore, this scoring method will eliminate the problems inherent in method 8, and will probably result in more stable ability estimates. In order to use this method, however, the problem of individuals for whom a clear basal or ceiling stratum cannot be determined must be solved.

10. The stradaptive test can be scored by determining the average difficulty of items answered correctly at the highest non-chance stratum.

This method is the average difficulty analogue of method 3. It essentially identifies the peaked test of highest difficulty which is not inappropriate for a given testee, eliminating those that are too difficult and those that are too easy. It should give ability estimates with good variability and fairly high stability.

The variety of scoring methods available suggests a number of interesting research possibilities using stradaptive tests. Scoring methods may vary in terms of psychometric characteristics, such as stability, shape of resulting score distributions, or correlations with scores on other testing strategies. Scoring methods may also vary in terms of validity and/or utility, with some methods better predicting external criteria or being more useful in different kinds of situations. Only future research, using a variety of empirical, simulation, and theoretical studies will determine which scoring methods are best suited for particular purposes.

## Consistency of Ability Estimates

The ten scoring methods described above, and others yet to be developed, all give "point estimates" of an individual's ability. Thus, they each return one value, based on some function of the difficulties of the items a testee has answered correctly, which indicates the point at which he falls on the underlying ability continuum. An analysis of the test records of individuals who have taken stradaptive tests shows additional information which reflects the consistency of the testee's response pattern. Such consistency data can be interpreted like data on the standard error of measurement; it indicates the range of confidence which can be attributed to a given ability point estimate. Individuals who are more consistent should have more stable ability estimates, while those who are less consistent should have less stable ability estimates. At present, this is only an hypothesis which will need empirical verification.

On stradaptive tests, individual differences occur in the number of strata between the basal stratum and the ceiling stratum. Thus, it is possible for some individuals to have the same score by one or more scoring methods (e.g., difficulty of the highest non-chance stratum), but the number of strata utilized in obtaining that score will differ widely. Some testees are consistent enough in their responses that their response records encompass only two or three strata. Other testees respond more inconsistently to the items, and their response records may encompass five or more strata between the basal and ceiling strata. Thus, the number of strata used by the testee can be a rough index of the consistency of his ability estimate, if items resulting from inappropriate entry points are eliminated. A related index would be the difference in average difficulties between the ceiling and basal strata.

A more meaningful consistency index might be the variance or standard deviation of the difficulties of the items answered correctly between the testee's basal and ceiling strata. This index would reflect more accurately the consistency of an individual's stradaptive test performance. It has the further advantage of being within the control of the tester. Since the variance is a mean, adding more items at or near the mid-point of the distribution of correct responses will reduce the variance. Reduction of this variance consistency estimate will occur then, by administering additional items at an individual's estimated ability level; since these items will have little or no deviation from his ability, the variance will continue to reduce with additional items. Testing could then continue in this fashion until a desired "standard error of measurement" was reached. At the same time that the variance reduction occurs by administering additional items, indicating greater confidence in the abilility estimate, the ability estimate itself should stabilize due to the greater number of items administered.

Individuals differ also in the number of items necessary to reach a termination criterion. In over 350 stradaptive tests administered to college students, the median number of items required to reach .termination was 18; the shortest stradaptive test required only 9 items and the longest required 160 items. Individuals who required a larger number of items also utilized a larger number of strata. The number of items required for termination, therefore, is a rough indication of an individual's consistency of response. Only further research on the relationship of this additional individual differences variable with other consistency data and with other data external to the stradaptive testing procedure will determine its utility.

## Illustrative Results from Stradaptive Testing

The previous sections have described the essential characteristics of the stradaptive test. However, to understand the method more completely, it is helpful to see the results of its application with actual testees. The following figures are graphical illustrations of the response records of a number of college students who took stradaptive tests.[3] The 9-stratum item pool used consisted of 229 5-response choice vocabulary items; the structure of the item pool is shown in Table 1. Entry point information was the student's report of his/her GPA as shown in Figure 3. An "up-one/down-one" branching rule was used. Termination occurred when a stratum was identified at which

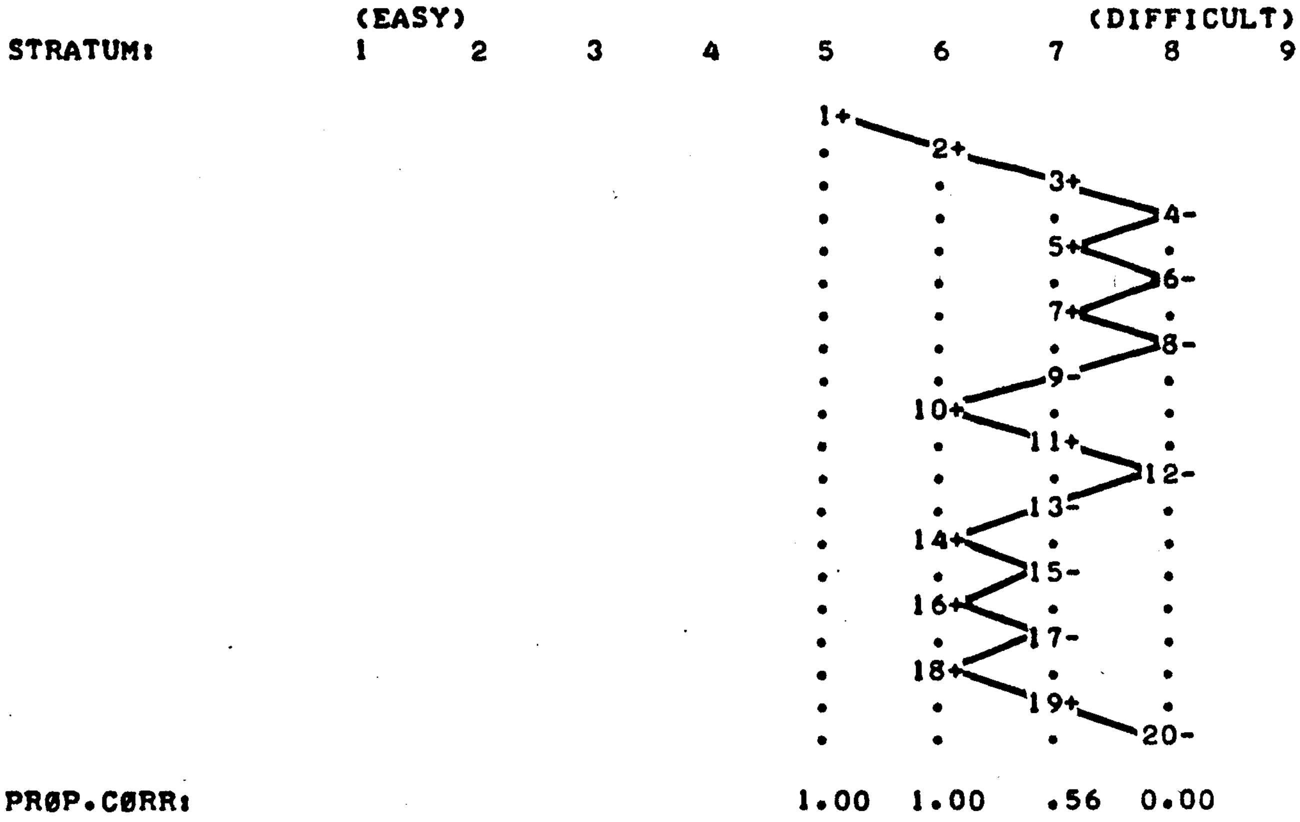
[3] The stradaptive test administration program was written by Robert Swisher; the display program was written by David Vale.

Figure 4

```
                    REPØRT ØN STRADAPTIVE TEST

NAME: WILLIAM W.                                DATE TESTED:  73/07/12

---------------------------------------------------------------------
```



```
                  TØTAL PRØPØRTIØN CØRRECT=  .550

        SCØRES ØN STRADAPTIVE TEST

        1. DIFFICULTY ØF MØST DIFFICULT ITEM CØRRECT=  1.49

        2. DIFFICULTY ØF THE N+1 TH ITEM=  1.44

        3. DIFFICULTY ØF HIGHEST NØN-CHANCE ITEM CØRRECT=  1.49

        4. DIFFICULTY ØF HIGHEST STRATUM
           WITH A CØRRECT ANSWER=  1.33

        5. DIFFICULTY ØF THE N+1 TH STRATUM=  1.33

        6. DIFFICULTY ØF HIGHEST NØN-CHANCE STRATUM=  1.33

        7. INTERPØLATED STRATUM DIFFICULTY=  1.37

        8. MEAN DIFFICULTY ØF ALL CØRRECT ITEMS=   .88

        9. MEAN DIFFICULTY ØF CØRRECT ITEMS BETWEEN
           CEILING AND BASAL STRATA                 =  1.28

       10. MEAN DIFFICULTY ØF ITEMS CØRRECT
           AT HIGHEST NØN-CHANCE STRATUM=  1.28
```

Table 2

Number of items administered (N) and cumulative proportion correct (p) by stage, for William W.

| | Stratum | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | | | 5 | | 6 | | 7 | | 8 | | 9 | | Total | |
| Stage | N | p | ... | N | p | N | p | N | p | N | p | N | p | N | p |
| 1 | | | | 1 | 1.00 | | | | | | | | | 1 | 1.00 |
| 2 | | | | | | 1 | 1.00 | | | | | | | 2 | 1.00 |
| 3 | | | | | | | | 1 | 1.00 | | | | | 3 | 1.00 |
| 4 | | | | | | | | | | 1 | 0.00 | | | 4 | .75 |
| 5 | | | | | | | | 2 | 1.00 | | | | | 5 | .80 |
| 6 | | | | | | | | | | 2 | 0.00 | | | 6 | .67 |
| 7 | | | | | | | | 3 | 1.00 | | | | | 7 | .71 |
| 8 | | | | | | | | | | 3 | 0.00 | | | 8 | .63 |
| 9 | | | | | | | | 4 | .75 | | | | | 9 | .56 |
| 10 | | | | | | 2 | 1.00 | | | | | | | 10 | .60 |
| 11 | | | | | | | | 5 | .80 | | | | | 11 | .64 |
| 12 | | | | | | | | | | 4 | 0.00 | | | 12 | .58 |
| 13 | | | | | | | | 6 | .67 | | | | | 13 | .54 |
| 14 | | | | | | 3 | 1.00 | | | | | | | 14 | .57 |
| 15 | | | | | | | | 7 | .57 | | | | | 15 | .53 |
| 16 | | | | | | 4 | 1.00 | | | | | | | 16 | .56 |
| 17 | | | | | | | | 8 | .50 | | | | | 17 | .53 |
| 18 | | | | | | 5 | 1.00 | | | | | | | 18 | .56 |
| 19 | | | | | | | | 9 | .56 | | | | | 19 | .58 |
| 20 | | | | | | | | | | 5 | 0.00 | | | 20 | .55 |

the proportion of correct responses was .20 or less, based on a minimum of five items completed at that stratum. Test items were presented to the student on a cathode-ray-terminal (CRT) with responses recorded through the CRT typewriter keyboard.

A typical response record. Figure 4 shows the stradaptive test performance of "William W.", a college sophomore. This test record is typical of the stradaptive test performance of college students. William was first presented with an entry point screen (Figure 3) and indicated that his cumulative grade point average to date was between 2.76 and 3.00. He thus began the stradaptive test at stratum 5. His answer to the first item was correct (indicated by a "+" in Figure 4), which branched him to the first available item in stratum 6. Correct answers to the second and third items resulted in his moving to stratum 8, where he received the first item from that more difficult peaked test. Since the stage 4 item was too difficult for him, his response was incorrect (-), and he branched downward to the first item in stratum 7. William then alternated between correct and incorrect responses for the items at stages 6 through 8, followed by an incorrect response to the stage 9 item. This returned him to stratum 6 for his tenth item. With a few minor deviations, William then essentially alternated between correct and incorrect responses from stages 11 through 20. Item 20 terminated the stradaptive test since the testing procedure had, at that point, located William's ceiling stratum; at stratum 8 William had answered all 5 items incorrectly.

Table 2 shows a complete "accounting" of William's stradaptive test performance. As the data in Table 2 indicate, tentative estimates of William's "basal" and "ceiling" strata were evident by stage 10; at that point he had 100% of the items correct at stratum 6, 75% correct at stratum 7 and none correct at stratum 8; his total percent correct at stage 10 was 60%. However, these percentages were based on only 2, 4, and 3 items respectively and therefore were not likely to be very stable. Since the termination criterion had not been met (i.e., 20% or less items correct based on 5 items administered at a stratum) the stradaptive test continued. As additional items were administered, William continued to answer all items at stratum 6 correctly, and at stratum 7 answered some items correctly and some incorrectly. By stage 19, he had completed the first 9 items available at stratum 7 and had answered 56% of those correctly. The final item administered (stage 20) was the fifth item at stratum 8, which he answered incorrectly.

The last column of Table 2 shows the proportion correct at each stage of the stradaptive test. That proportion shows a steady step-like decrease from 100% correct at stage 1 to 55% correct at stage 20. It is typical of stradaptive test performance for the proportion correct at the final stage to be near .50; in William's test performance the proportion correct stayed between .50 and .60 from stage 2 through termination.

Figure 4 also shows stradaptive test scores for William, using the scoring methods described earlier. As might be expected, the "highest difficulty" scores produced the highest ability estimates, and methods 1 and 3 gave the same results since William answered no items correctly at or above his ceiling stratum. Methods 4, 5 and 6 gave identical results for similar reasons; with a different set of test responses, however, these results would differ. The "average difficulty" methods gave the lowest ability estimates as a group, since the averages were lowered by the inclusion of the less difficult items.

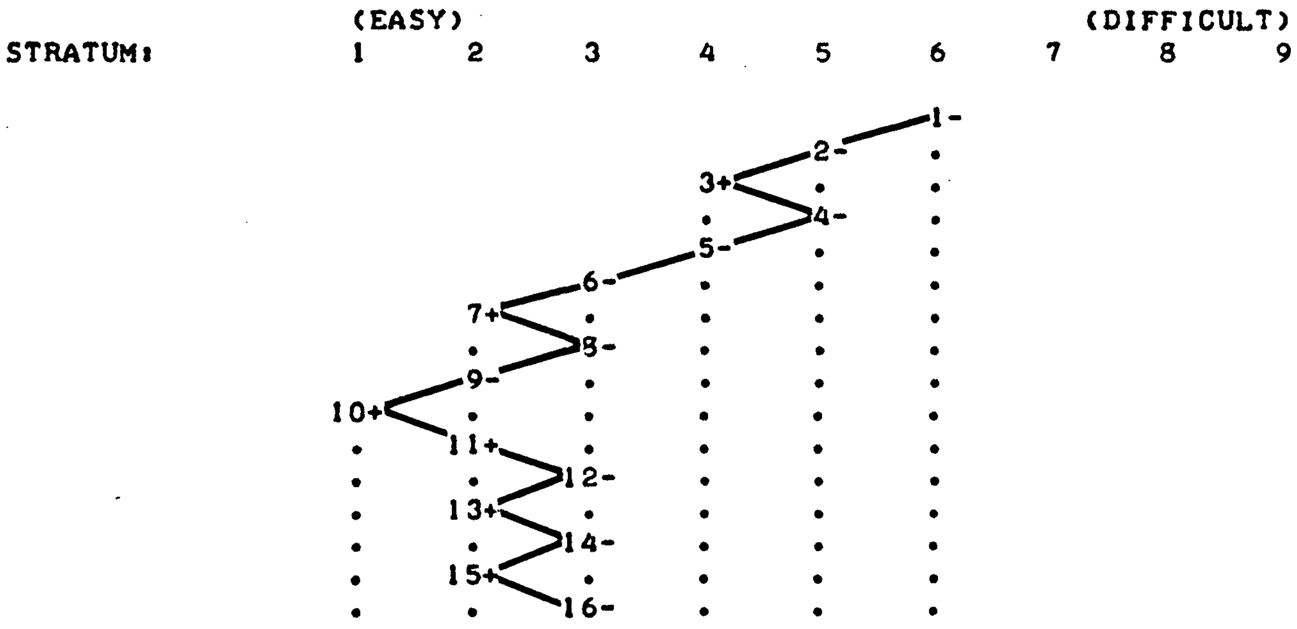
William's stradaptive test performance (Figure 4) is an example of a slightly low entry point. Because he entered at stratum 5, which was below his basal stratum 6, his response to the first item conveyed no information. However, it did serve to route him to the higher strata where testing was concentrated. Eliminating the first item administered from total proportion correct gives a proportion of .45 correct for William at the termination of testing.

High entry point. Occasionally an entry point is too high; an example is shown in Figure 5 for "Carol C." Carol reported her GPA to be in category 4, 3.01 to 3.25 (see Figure 3); this led to an entry at stratum 6. Her item responses quickly showed that the tests at strata 6, 5, 4, and 3 were too difficult for her. On the first six items Carol gave only one correct answer, an apparent "lucky guess" to a stratum 4 item. The routing procedure quickly brought Carol to strata 3, 2, and 1, which were composed of easier test items. Once she reached these strata her response pattern converged quickly on a region of the item pool in which she answered about 50% of the items correctly. Although her total proportion correct was only .375, eliminating the routing items due to the erroneous entry point (items 1 through 5), Carol obtained 5 correct answers out of 11 items in stages 6 through 10, for an effective proportion correct of .45. Disregarding the first 5 routing items, Carol's stradaptive test performance is similar to that of William's. In both cases the stradaptive test

Figure 5

REPØRT ØN STRADAPTIVE TEST

NAME: CARØL C. DATE TESTED: 73/07/12



PRØP.CØRR: 1.00 .80 0.00 .50 0.00 0.00

TØTAL PRØPØRTIØN CØRRECT= .375

SCØRES ØN STRADAPTIVE TEST

1. DIFFICULTY ØF MØST DIFFICULT ITEM CØRRECT= -.70

2. DIFFICULTY ØF THE N+1 TH ITEM= -1.92

3. DIFFICULTY ØF HIGHEST NØN-CHANCE ITEM CØRRECT= -1.68

4. DIFFICULTY ØF HIGHEST STRATUM
   WITH A CØRRECT ANSWER= -.63

5. DIFFICULTY ØF THE N+1 TH STRATUM= -1.92

6. DIFFICULTY ØF HIGHEST NØN-CHANCE STRATUM= -1.92

7. INTERPØLATED STRATUM DIFFICULTY= -1.73

8. MEAN DIFFICULTY ØF ALL CØRRECT ITEMS= -1.81

9. MEAN DIFFICULTY ØF CØRRECT ITEMS BETWEEN
   CEILING AND BASAL STRATA = -1.94

10. MEAN DIFFICULTY ØF ITEMS CØRRECT
    AT HIGHEST NØN-CHANCE STRATUM= -1.94

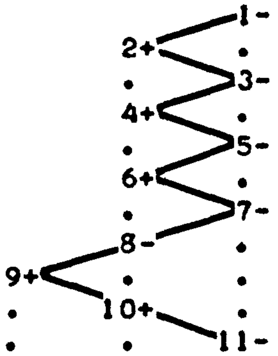
Figure 6

```
                    REPØRT ØN STRADAPTIVE TEST

NAME: JØHN J.                                   DATE TESTED:  73/04/09

--------------------------------------------------------------------

                  (EASY)                                    (DIFFICULT)
STRATUM:           1     2     3     4     5     6     7     8     9
```



```
PRØP.CØRR:                     1.00  .80  0.00

              TØTAL PRØPØRTIØN CØRRECT=  .455

        SCØRES ØN STRADAPTIVE TEST

        1. DIFFICULTY ØF MØST DIFFICULT ITEM CØRRECT=  -.52

        2. DIFFICULTY ØF THE N+1 TH ITEM=  -.75

        3. DIFFICULTY ØF HIGHEST NØN-CHANCE ITEM CØRRECT=  -.52

        4. DIFFICULTY ØF HIGHEST STRATUM
           WITH A CØRRECT ANSWER=  -.63

        5. DIFFICULTY ØF THE N+1 TH STRATUM=  -.63

        6. DIFFICULTY ØF HIGHEST NØN-CHANCE STRATUM=  -.63

        7. INTERPØLATED STRATUM DIFFICULTY=  -.44

        8. MEAN DIFFICULTY ØF ALL CØRRECT ITEMS=  -.81

        9. MEAN DIFFICULTY ØF CØRRECT ITEMS BETWEEN
           CEILING AND BASAL STRATA                      =  -.63

       10. MEAN DIFFICULTY ØF ITEMS CØRRECT
           AT HIGHEST NØN-CHANCE STRATUM=  -.63
```

identified a ceiling stratum (none correct or chance responding) a basal stratum (all correct), and a peaked test in between on which the testee obtained an intermediate proportion correct. In Carol's case the optimal peaked test was at stratum 2, on which she obtained 80% correct responses, while William's optimal peaked test was at stratum 7, on which he obtained 56% correct responses. It is interesting to note that William's entry point was lower than Carol's, yet their terminal ability levels were quite the reverse.

Rapid convergence. When the entry point estimate is accurate, the stradaptive test record can be quite short. Figure 6 shows an actual test record for "John J.". John entered at stratum 5 and immediately began alternating between correct and incorrect responses through stage 8. An incorrect response at stage 8 led to the identification of the basal stratum (although based on only one item) at stratum 3. Finally, an incorrect response on the stage 11 item permitted John to reach the termination criterion in only 11 items, having identified stratum 5 as John's ceiling stratum. John's ability level lies in the vicinity of stratum 4 at which he answered 80% of the items correctly. Over all 11 items administered, John answered 5, or a proportion of .455, correctly.

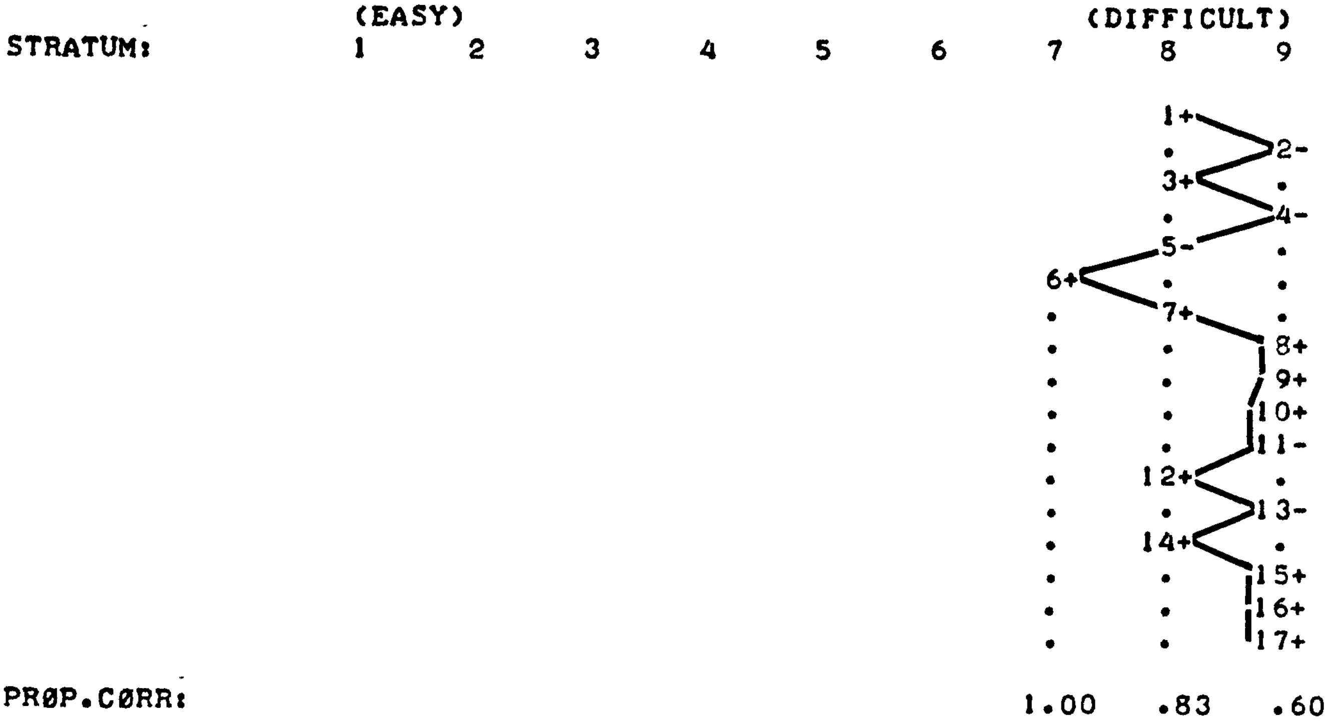
Item pool too easy. Occasionally the stradaptive item pool is too easy, or too difficult, for a testee. Figure 7 shows the stradaptive test performance of "Nancy N.". Nancy entered at stratum 8, based on a GPA estimate in the range of 3.51 to 3.75, almost an A average. With the exception of the stage 6 item, at stratum 7, testing of Nancy was confined to the difficult peaked tests at strata 8 and 9. Seventeen items were administered to Nancy, with 10 of them at stratum 9, the stratum with the most difficult items in the stradaptive item pool. Since stratum 9 contained only 10 items, testing was terminated. It is obvious that further testing of Nancy would be unproductive even if additional items were available at stratum 9. Nancy answered 83% of the items correctly at stratum 8, and 60% correctly at stratum 9. Since it would be quite unlikely that stratum 9 could be her ceiling stratum (.20 or less correct), no purpose would be served by further testing. In this case, the stradaptive test simply indicates that Nancy's ability is very high, but it is unable to give an estimate of exactly how high it is since she is apparently "off the top" of the most difficult test in the stradaptive pool. However, her ability is probably not as high as the individual who would answer all items correctly at stratum 9. The latter individual would answer 100% of the

Figure 7

REPØRT ØN STRADAPTIVE TEST

```
NAME: NANCY N.                                      DATE TESTED:  73/04/09

--------------------------------------------------------------------------
```



TØTAL PRØPØRTIØN CØRRECT= .706

SCØRES ØN STRADAPTIVE TEST

```
 1. DIFFICULTY ØF MØST DIFFICULT ITEM CØRRECT=  3.11

 2. DIFFICULTY ØF THE N+1 TH ITEM=      I

 3. DIFFICULTY ØF HIGHEST NØN-CHANCE ITEM CØRRECT=  3.11

 4. DIFFICULTY ØF HIGHEST STRATUM
    WITH A CØRRECT ANSWER=  2.62

 5. DIFFICULTY ØF THE N+1 TH STRATUM=  3.27

 6. DIFFICULTY ØF HIGHEST NØN-CHANCE STRATUM=  2.62

 7. INTERPØLATED STRATUM DIFFICULTY=  2.69

 8. MEAN DIFFICULTY ØF ALL CØRRECT ITEMS=  2.24

 9. MEAN DIFFICULTY ØF CØRRECT ITEMS BETWEEN
    CEILING AND BASAL STRATA                    =  2.35

10. MEAN DIFFICULTY ØF ITEMS CØRRECT
    AT HIGHEST NØN-CHANCE STRATUM=  2.63
```

items correctly, while Nancy answered only 60% correctly. Thus, the total proportion correct can be a rough indicator of the appropriateness of the stradaptive item pool for an individual. When that proportion, corrected for routing, is between .40 and .60, it indicates a test record appropriately adapted to the individual's ability level.

Two problems arose in computing scores for Nancy's stradaptive test performance. Scoring method 2, which determines score on the basis of the difficulty of the n+1th item could not be implemented for Nancy. Since she answered her last item correctly and it was the last item at stratum 9, the next item to be administered would have been an item at stratum 10. There were, however, only 9 strata in the stradaptive item pool. Thus, the difficulty of the n+1th item is indeterminate in Nancy's case, and an "I" is given on the computer report. A similar problem arose in computing the interpolated stratum difficulty score (method 7). Since Nancy answered 60% of the items correctly at stratum 9, her ability could be estimated to be above the mean difficulty of the stratum 9 peaked test ($z=2.62$, based on .50 correct). To compute the interpolated stratum difficulty score, the increment between the strata in the item pool, approximately .655, was added to the mean difficulty of stratum 9; Nancy's score was then interpolated into the interval between 2.62 and 3.27 by the formula given earlier.

Consistent vs. inconsistent response records. As indicated above, stradaptive test records can reflect individual differences in consistency of test performance. Figures 8 and 9 contrast the test records of "Tom T." and "Dixie D". In both cases entry into the item pool was at about the same level of difficulty; Tom entered at stratum 6 while Dixie began at stratum 7. For the first 8 items, both Tom and Dixie alternated between items at strata 6 and 7, and both had moved to the easier items at stratum 5 by the 10th stage of testing. After two items at stratum 5, Tom recovered quickly to stratum 6 and reached the termination criterion after 14 items. Tom's basal stratum was stratum 5, and stratum 7 was his ceiling stratum. His highest non-chance stratum was stratum 6, at which he answered 71% of the items correctly.

Dixie's test performance, although similar to Tom's in the earlier stages of testing, diverged sharply after the twelfth item. At that point she began to answer easier items incorrectly, finally being presented with an item from stratum 3 at the 17th stage of testing. Dixie's response

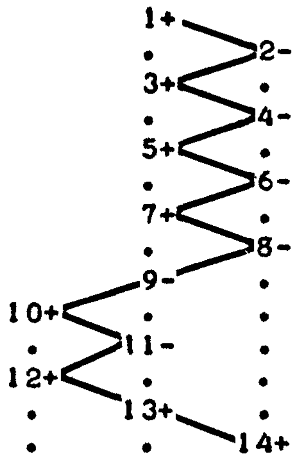
Figure 8

```
                    REPØRT ØN STRADAPTIVE TEST

NAME: TØM T.                                   DATE TESTED:  73/07/02

-----------------------------------------------------------------------

                 (EASY)                                     (DIFFICULT)
STRATUM:           1     2     3     4     5     6     7     8     9
```



```
PRØP.CØRR:                                    1.00   .71   .20

                   TØTAL PRØPØRTIØN CØRRECT=  .571
```

SCØRES ØN STRADAPTIVE TEST

1. DIFFICULTY ØF MØST DIFFICULT ITEM CØRRECT= 1.11

2. DIFFICULTY ØF THE N+1 TH ITEM= 1.89

3. DIFFICULTY ØF HIGHEST NØN-CHANCE ITEM CØRRECT= .79

4. DIFFICULTY ØF HIGHEST STRATUM
   WITH A CØRRECT ANSWER= 1.33

5. DIFFICULTY ØF THE N+1 TH STRATUM= 2.01

6. DIFFICULTY ØF HIGHEST NØN-CHANCE STRATUM= .65

7. INTERPØLATED STRATUM DIFFICULTY= .80

8. MEAN DIFFICULTY ØF ALL CØRRECT ITEMS= .52

9. MEAN DIFFICULTY ØF CØRRECT ITEMS BETWEEN
   CEILING AND BASAL STRATA = .59

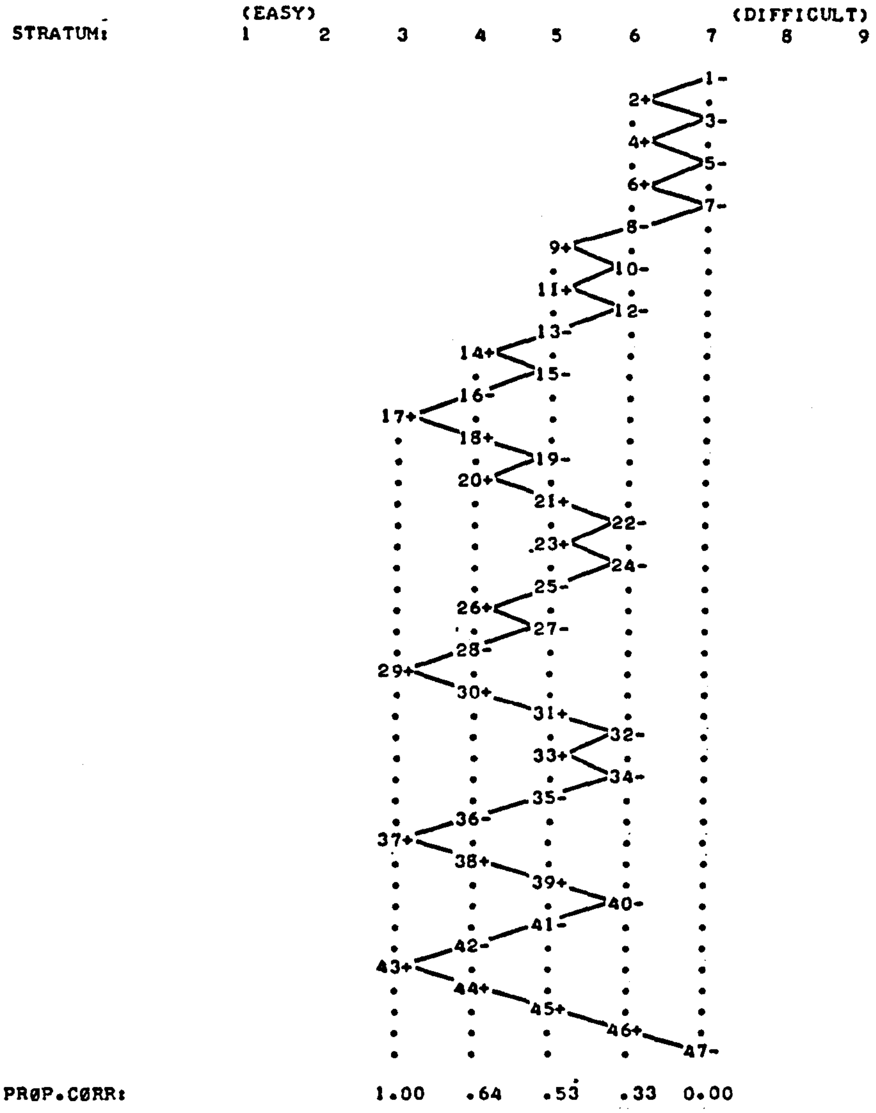
10. MEAN DIFFICULTY ØF ITEMS CØRRECT
    AT HIGHEST NØN-CHANCE STRATUM= .59

Figure 9

```
                    REPØRT ØN STRADAPTIVE TEST

NAME: DIXIE D.                                   DATE TESTED:  73/04/09

-----------------------------------------------------------------------
```



```
              TØTAL PRØPØRTIØN CØRRECT=  .489
      SCØRES ØN STRADAPTIVE TEST

      1. DIFFICULTY ØF MØST DIFFICULT ITEM CØRRECT=   .73

      2. DIFFICULTY ØF THE N+1 TH ITEM=   .78

      3. DIFFICULTY ØF HIGHEST NØN-CHANCE ITEM CØRRECT=   .73

      4. DIFFICULTY ØF HIGHEST STRATUM
         WITH A CØRRECT ANSWER=   .65

      5. DIFFICULTY ØF THE N+1 TH STRATUM=   .65

      6. DIFFICULTY ØF HIGHEST NØN-CHANCE STRATUM=   .65

      7. INTERPØLATED STRATUM DIFFICULTY=   .54

      8. MEAN DIFFICULTY ØF ALL CØRRECT ITEMS=  -.30

      9. MEAN DIFFICULTY ØF CØRRECT ITEMS BETWEEN
         CEILING AND BASAL STRATA                      =  -.09

     10. MEAN DIFFICULTY ØF ITEMS CØRRECT
         AT HIGHEST NØN-CHANCE STRATUM=   .59
```

record then shows a series of wide swings between items at stratum 3 and those at stratum 6. While many testees converge on strata that are contiguous, Dixie's responses seem to show a convergence somewhere between strata 3 and 6. Thus, ability estimates derived from Dixie's stradaptive testing are likely to be less precise than those from Tom's responses. Dixie finally worked her way back up to stratum 7 after 47 items to satisfy the termination criterion.

Dixie's testing thus used five of the available nine strata, while Tom used only three. For both Tom and Dixie the ceiling stratum was stratum 7, but while Tom's basal ability was at stratum 5, Dixie's was at stratum 3. Stratum 6 was the highest non-chance stratum for both, but Tom's ability is probably closer to that of stratum 7 than to stratum 5, since he answered 71% of the items correctly at stratum 6. Dixie's, however, is more toward stratum 5, since she answered only 33% correctly at stratum 6. The difference is reflected by the interpolated stratum difficulty scores of .80 and .54 for the two testees, respectively. These two response records show how stradaptive test performance can differ in terms of both number of items administered and the number of strata used for ability determination.

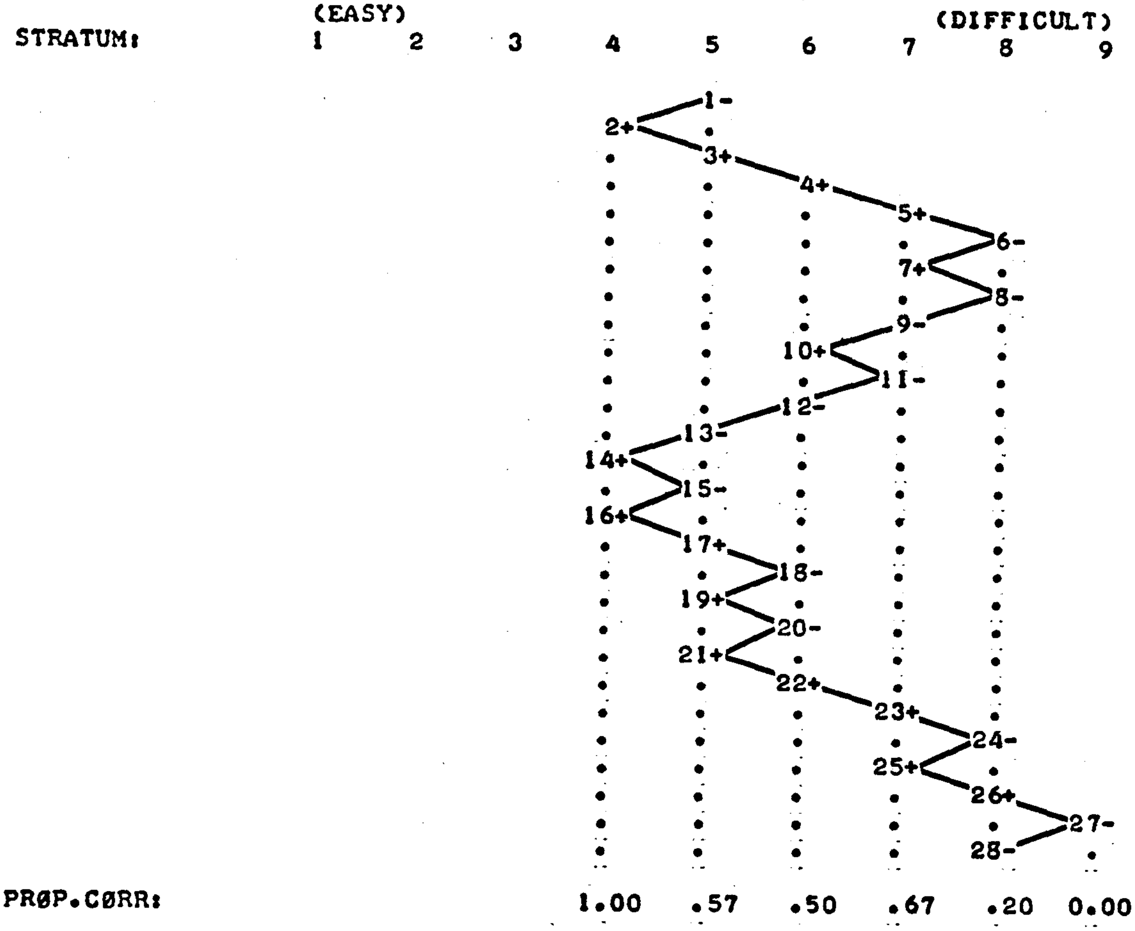
Another example of inconsistent stradaptive test performance is shown in Figure 10. This test record, for "Carl C.", shows a range of fluctuation even wider than that of Dixie D. (Figure 9). Carl seemed to answer almost optimally (i.e., about 50% correct) on the three peaked tests of strata 5, 6, and 7. His performance fluctuated rather consistently from strata 4 through 8, and he even attempted one item (27) at stratum 9, following a probable lucky guess at stratum 8. Carl's basal stratum was stratum 4(100% correct) and his ceiling stratum was stratum 8 (20% correct). Between these two he answered slightly more than 50% of the items correctly, with an overall proportion correct of .54. Carl's inconsistent performance on the stradaptive test stands in sharp contrast to that of, say, John J. (Figure 6), whose very consistent response record covered only three strata, and who reached the termination criterion in only 11 items. The utility of this information on individual differences in consistency of performance on the stradaptive test will be determined only through further research. Logically, however, it seems that such information could be used to derive individualized "standard errors of measurement."

## Implications of Proportion Correct Data

The data in Figures 4 through 10 illustrate an interesting characteristic of stradaptive test records. For

Figure 10

```
                    REPØRT ØN STRADAPTIVE TEST
NAME: CARL C.
                                            DATE TESTED:   73/07/12
--------------------------------------------------------------------
```



```
          TØTAL PRØPØRTIØN CØRRECT=  .536

      SCØRES ØN STRADAPTIVE TEST

      1. DIFFICULTY ØF MØST DIFFICULT ITEM CØRRECT=  2.31

      2. DIFFICULTY ØF THE N+1 TH ITEM=  1.17

      3. DIFFICULTY ØF HIGHEST NØN-CHANCE ITEM CØRRECT=  1.49

      4. DIFFICULTY ØF HIGHEST STRATUM
         WITH A CØRRECT ANSWER=  2.01

      5. DIFFICULTY ØF THE N+1 TH STRATUM=  1.33

      6. DIFFICULTY ØF HIGHEST NØN-CHANCE STRATUM=  1.33

      7. INTERPØLATED STRATUM DIFFICULTY=  1.44

      8. MEAN DIFFICULTY ØF ALL CØRRECT ITEMS=   .47

      9. MEAN DIFFICULTY ØF CØRRECT ITEMS BETWEEN
         CEILING AND BASAL STRATA                 =   .60

     10. MEAN DIFFICULTY ØF ITEMS CØRRECT
         AT HIGHEST NØN-CHANCE STRATUM=  1.27
```

most individuals completing a stradaptive test, the proportion of correct responses at the various strata decreases as the difficulty of the strata increases. These results are summarized in Figure 11, which plots the proportion of correct responses at each stratum. With the exception of the plots for Carl C. and Carol C., these plots resemble item trace lines (Lord & Novick, 1968). The steepness of the slope can be interpreted as an index of the consistency of responses of the individual and the capability of the item pool to "discriminate" that individual's ability level. The point of inflection of the curve (i.e., the point on the horizontal axis at which the testee answers 50% of the items correctly) could be interpreted as the "difficulty" of the item pool for the individual, or his position on the latent ability continuum.

Reasoning analogically from item characteristic curve theory, non-regular item characteristic curves, such as those for Carl C. and Carol C., might indicate item pool-testee interactions which are inappropriate. Thus, both Carol and Carl might not be interacting with the item pool on a unidimensional continuum. In order to get a more accurate ability estimate for such testees, it might be necessary to multidimensionally scale their response patterns to obtain subsets of test items (if possible) on which they responded in unidimensional fashion, as indicated by their test response "trace lines." Thus, Carl and Carol's response records might be analyzed by appropriate scaling methods to find the intra-individual probabilistic Guttman-type scales underlying their response patterns.

The "trace line" plots for John J., Tom T. and William W. approximate the classic step function Guttman-type trace line. Dixie D.'s trace line plot is very similar to the normal ogive probabilistic analogue of the Guttman trace line. Future research based on stradaptive tests with a large number of strata may lead to mathematization of these trace line ideas, which in turn may lead to greater utility for this type of test data.

It is interesting to note that the stradaptive test performance of many testees results in a Guttman-like scaling of the testee's performance with respect to the item pool. Since the stradaptive test developed from the testing rationale originally proposed by Binet, it follows that perhaps Binet's ability testing logic had embedded in it an unarticulated primitive version of Guttman's ideas and the present-day derivates of modern test theory as derived from latent trait theory.

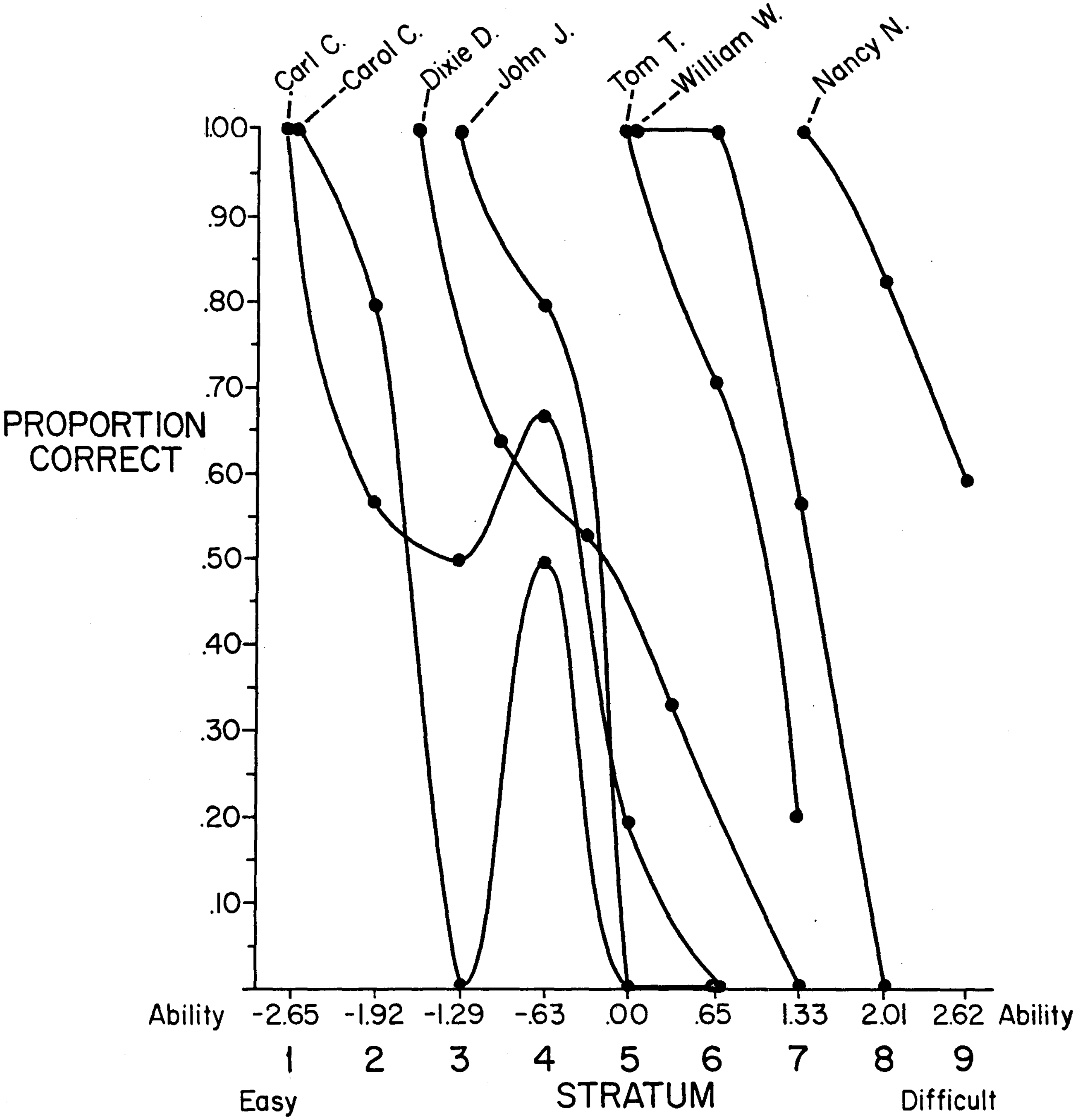


Figure 11. Proportion correct at each stratum, by individual

## Conclusions

The stradaptive test is an operational computer-based testing model which draws simultaneously from Binet's pioneering work in ability measurement and from ideas in modern test theory. The testing procedure makes no restrictive assumptions about the nature of underlying ability distributions (beyond those involved in norming the item pool), and its implementation does not require complicated mathematical calculations. The procedure is also flexible with respect to size and composition of the item pool, branching rules, termination rules, and scoring methods. Data derived from the stradaptive test response record, including number of items completed, range of difficulties used, patterns of movement through the item pool, and various other methods of measuring a testee's interaction with a specified item pool appear to have promise as new sources of information derivable from ability testing.

The availability of the stradaptive testing strategy poses many new research questions. Among these are the optimal characteristics (e.g., size, number of strata) of the stradaptive item pool, methods of selecting and placing items in the pool, variations in branching rules, applications of stochastic models to the branching process, variations in step size, effects of various termination rules, the reliability and utility of the various scoring methods proposed and those yet to be developed, methods of expressing an individual's consistency or the accuracy of test scores, methods of controlling the accuracy of test scores within the stradaptive framework, and relationships of stradaptive scores and ability estimates to those derived from other adaptive strategies. These research questions should be studied by a variety of approaches, including live testing empirical studies, simulation studies, and theoretical studies, with the results of each approach supporting and nourishing research using the other approaches.

## References

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.

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.

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

Cronbach, L. G., Gleser, G. C., Nanda H. & Rajaratnam, N. The dependability of behavioral measurements: theory of generalizability for scores and profiles. New York: Wiley, 1972.

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

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

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

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

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

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

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

Terman, L. M. & Merrill, M. A. Stanford-Binet Intelligence Scale. Boston: Houghton Mifflin, 1960.

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

Weiss, D. J. Strategies of computerized ability testing. Research Report 73-x, Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, 1973 (in preparation).

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

# DISTRIBUTION LIST

## Navy

4 Dr. Marshall J. Farr, Director
Personnel & Training Research Programs
Office of Naval Research
Arlington, VA 22217

1 Director
ONR Branch Office
495 Summer Street
Boston, MA 02210
ATTN: C. M. Harsh

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

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

1 Office of Naval Research
Area Office
207 West 24th Street
New York, NY 10011

6 Director
Naval Research Laboratory
Code 2627
Washington, DC 20390

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

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

1 Chief of Naval Technical Training
Naval Air Station Memphis (75)
Millington, TN 38054
ATTN: Dr. G. D. Mayo

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

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

1 Commander
Naval Air Reserve
Naval Air Station
Glenview, IL 60026

1 Commander
Naval Air Systems Command
Department of the Navy
AIR-413C
Washington, DC 20360

1 Mr. Lee Miller (AIR 413E)
Naval Air Systems Command
5600 Columbia Pike
Falls Church, VA 22042

1 Dr. Harold Booher
NAVAIR 415C
Naval Air Systems Command
5600 Columbia Pike
Falls Church, VA 22042

1 CAPT John F. Riley, USN
Commanding Officer
U.S. Naval Amphibious School
Coronado, CA 92155

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

1 Dr. Richard J. Niehaus
Office of Civilian Manpower Management
Code 06A
Department of the Navy
Washington, DC 20390

1 CDR Richard L. Martin, USN
COMFAIRMIRAMAR F-14
NAS Miramar, CA 92145

1 Research Director, Code 06
Research and Evaluation Department
U.S. Naval Examining Center
Great Lakes, IL 60088
ATTN: C. S. Winiewicz

1 Chief
Bureau of Medicine and Surgery
Code 413
Washington, DC 20372

1 Program Coordinator
Bureau of Medicine and Surgery (Code 71G)
Department of the Navy
Washington, DC 20372

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

1 Technical Reference Library
Naval Medical Research Institute
National Naval Medical Center
Bethesda, MD 20014

1 Chief
Bureau of Medicine and Surgery
Research Division (Code 713)
Department of the Navy
Washington, DC 20372

1 Dr. John J. Collins
Chief of Naval Operations (OP-987F)
Department of the Navy
Washington, DC 20350

1 Technical Library (Pers-11B)
Bureau of Naval Personnel
Department of the Navy
Washington, DC 20360

1 Head, Personnel Measurement Staff
Capital Area Personnel Office
Ballston Tower #2, Room 1204
801 N. Randolph Street
Arlington, VA 22203

1 Dr. James J. Regan, Technical Director
Navy Personnel Research
and Development Center
San Diego, CA 92152

1 Mr. E. P. Somer
Navy Personnel Research
and Development Center
San Diego, CA 92152

1 Dr. Norman Abrahams
Navy Personnel Research
and Development Center
San Diego, CA 92152

1 Dr. Bernard Rimland
Navy Personnel Research
and Development Center
San Diego, CA 92152

1 Commanding Officer
Navy Personnel Research
and Development Center
San Diego, CA 92152

1 Superintendent
Naval Postgraduate School
Monterey, CA 92940
ATTN: Library (Code 2124)

1 Mr. George N. Graine
Naval Ship Systems Command
(SHIPS 03H)
Department of the Navy
Washington, DC 20360

1 Technical Library
Naval Ship Systems Command
National Center, Building 3
Room 3S08
Washington, DC 20360

1 Commanding Officer
Service School Command
U.S. Naval Training Center
San Diego, CA 92133
ATTN: Code 303

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

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

1 CDR Fred Richardson
Navy Recruiting Command
BCT #3, Room 215
Washington, DC 20370

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

1 Dr. H. Wallace Sinaiko
c/o Office of Naval Research (Code 450)
Psychological Sciences Division
Arlington, VA 22217

1 Dr. Martin F. Wiskoff
Navy Personnel Research
and Development Center
San Diego, CA 92152

<u>Army</u>

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

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

1 Director of Research
U.S. Army Armor Human Research Unit
ATTN: Library
Building 2422 Morade Street
Fort Knox, KY 40121

1 U.S. Army Research Institute for the
Behavioral and Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

1 Commanding Officer
ATTN: LTC Montgomery
USACDC - PASA
Ft. Benjamin Harrison, IN 46249

1 Commandant
United States Army Infantry School
ATTN: ATSIN-H
Fort Benning, GA 31905

1 U.S. Army Research Institute
Commonwealth Building, Room 239
1300 Wilson Boulevard
Arlington, VA 22209
ATTN: Dr. R. Dusek

1 Mr. Edmund F. Fuchs
U.S. Army Research Institute
1300 Wilson Boulevard
Arlington, VA 22209

1 Commander
U.S. Theater Army Support Command,
Europe
ATTN: Asst. DCSPER (Education)
APO New York 09058

1 Dr. Stanley L. Cohen
Work Unit Area Leader
Organizational Development Work Unit
Army Research Institute for Behavioral
and Social Science
1300 Wilson Boulevard
Arlington, VA 22209

<u>Air Force</u>

1 Headquarters, U.S. Air Force
Chief, Personnel Research and Analysis
Division (AF/DPSY)
Washington, DC 20330

1 Research and Analysis Division
AF/DPXYR Room 4C200
Washington, DC 20330

1 AFHRL/AS (Dr. G. A. Eckstrand
Wright-Patterson AFB
Ohio 45433

1 AFHRL/MD
701 Prince Street
Room 200
Alexandria, VA 22314

1 Dr. Robert A. Bottenberg
AFHRL/PES
Lackland AFB, TX 78236

1 Personnel Research Division
AFHRL
Lackland Air Force Base
Texas 78236

1 AFOSR(NL)
1400 Wilson Boulevard
Arlington, VA 22209

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

1 CAPT Jack Thorpe, USAF
Department of Psychology
Bowling Green State University
Bowling Green, OH 43403

Marine Corps

1 Commandant, Marine Corps
Code A01M-2
Washington, DC 20380

1 COL George Caridakis
Director, Office of Manpower Utilization
Headquarters, Marine Corps (A01H)
MCB
Quantico, VA 22134

1 Dr. A. L. Slafkosky
Scientific Advisor (Code Ax)
Commandant of the Marine Corps
Washington, DC 20380

1 Mr. E. A. Dover
Manpower Measurement Unit (Code A01M-2)
Arlington Annex, Room 2413
Arlington, VA 20370

Coast Guard

1 Mr. Joseph J. Cowan, Chief
Psychological Research Branch (P-1)
U.S. Coast Guard Headquarters
400 Seventh Street, SW
Washington, DC 20590

Other DOD

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

1 Mr. Helga Yeich, Director
Program Management, Defense Advanced
Research Projects Agency
1400 Wilson Boulevard
Arlington, VA 22209

1 Dr. Ralph R. Canter
Director for Manpower Research
Office of Secretary of Defense
The Pentagon, Room 3C980
Washington, DC 20301

Other Government

1 Dr. Lorraine D. Eyde
Personnel Research and Development Center
U.S. Civil Service Commission, Room 3458
1900 E. Street, N.W.
Washington, DC 20415

1 Dr. Vern Urry
Personnel Research and Development
Center
U.S. Civil Service Commission
Washington, DC 20415

Miscellaneous

1 Dr. Scarvia Anderson
Executive Director for Special
Development
Educational Testing Service
Princeton, NJ 08540

1 Dr. Richard C. Atkinson
Stanford University
Department of Psychology
Stanford, CA 94305

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

1 Mr. H. Dean Brown
Stanford Research Institute
333 Ravenswood Avenue
Menlo Park, CA 94025

1 Mr. Michael W. Brown
Operations Research, Inc.
1400 Spring Street
Silver Spring, MD 20910

1 Dr. Ronald P. Carver
American Institutes for Research
8555 Sixteenth Street
Silver Spring, MD 20910

1 Century Research Corporation
4113 Lee Highway
Arlington, VA 22207

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

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

1 Dr. Norman R. Dixon
Associate Professor of Higher
Education
University of Pittsburgh
617 Cathedral of Learning
Pittsburgh, PA 15213

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

1 Dr. Marvin D. Dunnette
University of Minnesota
Department of Psychology
N492 Elliott Hall
Minneapolis, MN 55455

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

1 Dr. Edwin A. Fleishman
American Institutes for Research
8555 Sixteenth Street
Silver Spring, MD 20910

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

1 Dr. Albert S. Glickman
American Institutes for Research
8555 Sixteenth Street
Silver Spring, MD 20910

1 Dr. Duncan N. Hansen
Florida State University
Center for Computer-Assisted Instruction
Tallahassee, FL 32306

1 Dr. Harry H. Harman
Educational Testing Service
Division of Analytical Studies
and Services
Princeton, NJ 08540

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

1 Dr. M. D. Havron
Human Sciences Research, Inc.
Westgate Industrial Park
7710 Old Springhouse Road
McLean, VA 22101

1 Human Resources Research Organization
Division #3
P.O. Box 5787
Presidio of Monterey, CA 93940

1 Human Resources Research Organization
Division #4, Infantry
P.O. Box 2086
Fort Benning, GA 31905

1 Human Resources Research Organization
Division #5, Air Defense
P.O. Box 6057
Fort Bliss, TX 79916

1 Human Resources Research Organization
Division #6, Library
P.O. Box 428
Fort Rucker, AL 36360

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

1 Dr. Norman J. Johnson
Carnegie-Mellon University
School of Urban and Public Affairs
Pittsburgh, PA 15213

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

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

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

1 Mr. Edmond Marks
109 Grange Building
Pennsylvania State University
University Park, PA 16802

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

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

1 Dr. Robert D. Pritchard
Assistant Professor of Psychology
Purdue University
Lafayette, IN 47907

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

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

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

1 Dr. Benjamin Schneider
University of Maryland
Department of Psychology
College Park, MD 20742

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

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

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

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

1 Dr. Charles A. Ullmann
Director, Behavioral Sciences Studies
Information Concepts Incorporated
1701 No. Ft. Myer Drive
Arlington, VA 22209

1 Dr. H. Peter Dachler
University of Maryland
Department of Psychology
College Park, MD 20742

# AN EMPIRICAL STUDY OF COMPUTER-ADMINISTERED TWO-STAGE ABILITY TESTING

Nancy E. Betz

and

David J. Weiss

Research Report 73-4

Psychometric Methods Program
Department of Psychology
University of Minnesota

October 1973

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

Security Classification

## DOCUMENT CONTROL DATA - R & D

*(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)*

1. ORIGINATING ACTIVITY *(Corporate author)*

Department of Psychology
University of Minnesota

2a. REPORT SECURITY CLASSIFICATION

Unclassified

2b. GROUP

3. REPORT TITLE

An Empirical Study of Computer-Administered Two-Stage Ability Testing

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

Technical Report

5. AUTHOR(S) *(First name, middle initial, last name)*

Nancy E. Betz and David J. Weiss

| | | |
|---|---|---|
| 6. REPORT DATE<br>October 1973 | 7a. TOTAL NO. OF PAGES<br>49 | 7b. NO. OF REFS<br>35 |

8a. CONTRACT OR GRANT NO.

N00014-67-A-0113-0029

b. PROJECT NO.

NR150-343

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)

Research Report 73-4
Psychometric Methods Program

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*

10. DISTRIBUTION STATEMENT

Approved for public release; distribution unlimited

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY

Personnel & Training Research Programs
Office of Naval Research

13. ABSTRACT

A two-stage adaptive test and a conventional peaked test were constructed and administered on a time-shared computer system to students in undergraduate psychology courses. Comparison of the score distributions yielded by the two tests showed that the two-stage test scores were somewhat more variable than the linear test scores, and that the distribution of two-stage scores was normal, whereas that of the linear test scores tended toward flatness. The two-stage test had higher test-retest stability than the conventional when the effect of memory of the items was taken into account. The relationship between the two-stage and conventional test scores was relatively high and primarily linear but left about 20% of the reliable variance in the conventional test scores unaccounted for. Further analyses of the two-stage test showed that the difficulty levels of the measurement tests were not optimal, and that 4 to 5% of the testees were misclassified into measurement tests. The relatively poor internal consistency of the measurement tests in comparison to that of the routing test and the conventional test was apparently due to the extreme homogeneity of ability within the measurement test sub-groups. The findings of the study were interpreted as favorable to continued exploration of two-stage testing procedures. Suggestions for possible ways to improve the characteristics of the two-stage testing strategy are offered.

| 14. KEY WORDS | LINK A | | LINK B | | LINK C | |
|---|---|---|---|---|---|---|
| | ROLE | WT | ROLE | WT | ROLE | WT |
| testing | | | | | | |
| ability testing | | | | | | |
| two-stage testing | | | | | | |
| computerized testing | | | | | | |
| adaptive testing | | | | | | |
| branched testing | | | | | | |
| individualized testing | | | | | | |
| tailored testing | | | | | | |
| programmed testing | | | | | | |
| response-contingent testing | | | | | | |
| automated testing | | | | | | |

# Contents

# AN EMPIRICAL STUDY OF COMPUTER-ADMINISTERED TWO-STAGE ABILITY TESTING

The growth and refinement of time-shared computer facilities has made it feasible to consider new approaches to the measurement of abilities. One such approach involves varying test item presentation procedures according to the characteristics of the individual being tested; this approach has been referred to as sequential testing (Cronbach and Gleser, 1957; Evans, 1953; Krathwohl and Huyser, 1956; Paterson, 1962), branched testing (Bayroff, 1964), programmed testing (Cleary, Linn, and Rock, 1968a), individualized measurement (Weiss, 1969), tailored testing (Lord, 1970), response-contingent measurement (Wood, 1971, 1973), and, most recently, adaptive testing (Weiss and Betz, 1973).

One model of adaptive testing is the two-stage procedure. This testing strategy consists of a routing test followed by one of a series of second-stage or "measurement" tests, each of which consists of items concentrated at a different level of difficulty. The purpose of the routing test is to give an initial estimate of an individual's ability so that he may be routed to the measurement test most appropriate to his ability. Cronbach & Gleser (1957) appear to have been the first to suggest the use of two-stage testing procedures. Weiss (1973) describes several variations of the basic two-stage strategy and compares them with other strategies of adaptive ability testing.

The first reported study of the two-stage procedure was an empirical study by Angoff and Huddleston (1958). They compared two-stage procedures with conventional "broad range" ability tests of verbal and mathematical abilities from the College Entrance Examination Board's Scholastic Aptitude Test. The two-stage test measuring verbal abilities consisted of a 40-item routing test and two 36-item measurement tests; their two-stage mathematical abilities test consisted of a 30-item routing test and two 17-item measurement tests. Nearly 6,000 students from 19 different colleges were tested, and all testing was timed. In the procedure followed, routing did not actually occur (i.e., the routing test was not scored prior to the administration of the measurement tests); rather, tests were administered in sufficient combinations to allow a determination of the effects of actual routing, had it occurred.

Results showed the measurement tests to be more reliable in the groups for which they were intended than conventional broad-range tests. Predictive validities of the measurement tests, using grade point averages as the criterion, were slightly higher than those of the conventional tests. Their data also showed, however, that about 20% of the testees would have been misclassified, or routed to an inappropriate measurement test.

A series of studies of two-stage procedures was reported by Cleary, Linn, and Rock (1968a, b; Linn, Rock, and Cleary, 1969). These were "real data" simulation studies, using the responses of 4,885 students to the 190 verbal items of the School and College Aptitude Tests and the Sequential Tests of Educational Progress. The total group was randomly split into a development group and a cross-validation group. Four 20-item measurement tests were constructed by dividing the total score distribution on the "parent" test into quartiles and finding the 20 items which had the highest within-quartile point-biserial correlations with the total test score.

Cleary _et al_. studied four different procedures of routing individuals to the measurement tests. The "broad-range" routing procedures consisted of a 20-item routing test with a rectangular distribution of item difficulties. Based on their scores on these 20-items, individuals were routed into one of the four measurement tests. The second strategy was a double-routing or two-phase procedure. In the first phase, scores on 10 items of median difficulty (p=.5) were used to divide the group into halves. The second phase used two additional 10-item routing tests; scores on these sets of 10 items were used to divide each first-phase subgroup into halves, yielding a total of four groups. The third routing procedure, called the "group discrimination" procedure, used the 20 items with the largest between-quartile differences in item difficulties.

The fourth procedure, called "sequential" routing, utilized the framework of the sequential sampling procedures developed by the Statistical Research Group (1945) and Wald (1950) and a specific procedure developed by Armitage (1950). In this method items would be administered to subjects one at a time. After scoring each item, "likelihood ratios" were computed and a decision was made either to assign the examinee to one of the four measurement tests or to administer another item. If the examinee had not been classified after all 23 routing items were administered, he was assigned to the group yielding the largest likelihood ratio. Cleary _et al_. also used a 3-group sequential procedure with a maximum of 20 routing items.

Scores on the two-stage tests were initially determined by scaling the measurement tests using linear regression weights to predict the total score on the parent test. A later study (Linn et al., 1969) added the routing score information to the scaled measurement test score.

Correlations between the two-stage test scores (based on a maximum of 43 items) and scores on the 190-item parent test were almost as high as the reliability estimates of the parent test. Scores from the sequential routing procedure correlated highest with total score, followed by 40- and 42-item conventional tests, the group discrimination, broad range, and double-routing procedures. Since the best short conventional test was found to require about 35% more items to achieve the same level of accuracy as the 3-group sequential procedure, it was concluded that two-stage tests can permit large reductions in the number of items administered to an individual with little or no loss in accuracy.

Validity results, in terms of correlations with external criteria of scores on the College Entrance Examination Board Tests and the Preliminary Scholastic Aptitude Tests, were even more favorable for the two-stage tests than were correlations with total test score. The group discrimination and 3-group sequential procedures yielded the highest correlations with the criteria. With the exception of the double-routing strategy, all of the two-stage procedures had higher validities than conventional tests of equivalent lengths. In most cases, the 40-item two-stage tests had higher validities than 50-item conventional tests, and in five comparisons they had higher validities than did the 190-item parent test. Thus, it was demonstrated that two-stage tests can achieve high predictive accuracy with substantially fewer items than would be necessary in a conventional test, although the data of Cleary et al., like that of Angoff and Huddleston, showed a misclassification rate of about 20%.

Lord (1971d) presents results from theoretical studies of two-stage testing procedures. All of his analyses were based on the mathematics of item characteristic curve theory and the following assumptions: 1) a fixed number of items administered to each examinee, 2) dichotomous (right-wrong) scoring, 3) normal ogive item characteristic curves, 4) a unidimensional set of items, 5) all items of equal discriminations, 6) peaked routing and measurement tests (i.e., all items in each subtest were of the same difficulty), and 7) linear (i.e., non-branched) routing and measurement tests. Lord studied about 200 different strategies, varying the total number of items (15 or 60), the number of alternative measurement tests, the cutting points for assignment to the second-stage tests, methods of scoring both the routing test and the entire two-stage procedure, and whether or not random guessing was assumed (for a 5-choice item, within the 60-item tests only). Lord compared each two-stage strategy with a peaked conventional test of equivalent length in terms of information functions, which indicate the relative numbers of items required to achieve equivalent precision of measurement. Precision can be defined as the capability of responses to a set of test items to accurately represent the "true ability" of hypothetical individuals.

Lord found that the linear test provided better measurement around the mean ability level of the group, but that the two-stage procedures provided increasingly better measurement with increased divergence from the mean ability level. The finding that the peaked linear test provided better measurement around the mean ability level has been supported by Lord's other theoretical studies comparing peaked ability tests with tests "administered" under a variety of adaptive testing strategies (Lord, 1970, 1971a, 1971c); thus, the peaked test always provided more precise measurement than the adaptive test when ability was at the point at which the test was peaked. However, as an individual's ability deviated from the average, the peaked test provided less precise measurement, and the adaptive test provided more precise measurement.

The importance of these findings is that they indicate that the most precise or accurate measurement for any individual will be obtained by administering to him/her a test peaked at a difficulty level equal to that individual's ability level. Thus, test items should be of median, or p=.50, difficulty for each individual, rather than of median difficulty for a group of individuals varying in ability.

But ability level, and thus the appropriate level of item difficulty for an individual, is not usually known in advance; it is the test's function to measure it. The two-stage strategy provides one method of adapting the difficulty of the test to the individual's ability level, in an effort to achieve more precise measurement. The routing test gives an initial estimate of an individual's ability level, and he/she is then routed or assigned to that "measurement" test which is peaked at a difficulty level close to his estimated ability.

Lord's theoretical study of two-stage testing procedures, based on the notion that a short routing test can be used to find the optimal peaked measurement test for any given individual, as well as the studies of Angoff and Huddleston (1958) and Cleary _et al_. (1968a,b; Linn _et al_., 1969) show considerable potential for two-stage tests, in terms of increases in internal consistency reliability, validity, and precision of measurement. However, only Angoff and Huddleston's was an empirical study, and even this study was not able to account for the effects of actual routing. The purpose of the present study, then, was to begin an empirical evaluation of two-stage testing procedures; the study involved the development, computer-controlled administration, and comparison of a two-stage test and a peaked conventional test.

## METHOD

### Design

This study was part of a larger program of research involving a series of empirical comparisons of a number of major strategies of adaptive testing. These studies were directed at answering two major questions: 1) Does adaptive testing show any advantages as compared to conventional ability testing procedures? and 2) Are some strategies of adaptive testing superior to others? To answer these questions, the studies were designed to permit the investigation of 1) the psychometric characteristics of tests administered under each adaptive strategy, in comparison with conventional linear tests, 2) the test-retest stability of ability estimates derived from each strategy, 3) the relationships between ability estimates derived from different adaptive strategies, and 4) the relationship between ability estimates derived from conventional testing and each of the adaptive strategies.

The design involved the construction and computer-controlled administration of tests using each adaptive strategy and a conventional linear test. So that data concerning the inter-relationships between strategies could be obtained, the tests were administered in pairs such that each combination of two tests would be administered to a large group of subjects. To obtain test-retest stability data, tests were re-administered to the same individuals after an interval of about six weeks.

In the first phase of the research, a two-stage, a flexilevel (Lord, 1971b), and a conventional linear test were constructed. Each test consisted of 40 items drawn from a common item pool but selected so that there would be no overlapping of items between tests. The tests were then administered two at a time to a total group of about

350 individuals such that each combination of two tests was given to about 100 individuals.

To examine the possibility of fatigue or practice effects or an interaction between test sequence and testing strategy, the order of administration of the tests within each combination was randomized on the first testing so that each test would be administered first to approximately half the testees and administered second to the other half. Retests were administered in the same order as the subject had initially received them.

Computer administration was necessary only for the adaptive tests, but the conventional linear test was also computer-administered to control for the possibility of "novelty" effects resulting from an atypical mode of test administration.

Although the first phase included the administration of a flexilevel test, the results of its administration will be reported in a later paper. The present paper is concerned only with the evaluation and comparison of the characteristics of the two-stage and the linear test and with the relationship between ability estimates derived from the two tests.

Of interest, first of all, were the characteristics of the score distributions yielded by the tests. It was expected that the two-stage test, because it adapts the difficulties of the items to the ability level of the testee, would utilize more of the available score distribution than would the conventional test. On a conventional "peaked" test, item difficulties are appropriate for individuals of average ability but may be inappropriate for testees who deviate from the average ability at which the test is peaked. Scores of high ability individuals may be artificially depressed if the items are too easy for them, and scores of low ability subjects may be artificially inflated if they correctly guess the answers to the large number of items that will be too difficult for them. In the two-stage test, however, high ability subjects would be routed to more difficult measurement tests, thus giving more "top" to the test, and low ability subjects would take measurement items more appropriate to their ability level, thus reducing the effects of random guessing. That the probability of random guessing decreases as item difficulties get closer to the subject's ability level has been suggested by Lord (1970), Owen (1969), Urry (1970), and Wood (1971), among others. Thus, because the two-stage test adapts item difficulties to the testee's ability level, two-stage test scores should have higher variability than scores from peaked conventional tests. In addition, the score distributions were examined to determine whether the tests yielded skewed, rectangular, peaked, or non-unimodal distributions.

Another psychometric consideration was the internal consistency reliability of the tests. The purpose of the routing test is to assign each individual to that measurement test composed of items most appropriate for him. Thus, routing, if it is effective, should form subgroups of individuals for whom the assigned measurement test is composed of items of appropriate difficulty. For 5-alternative multiple-choice items, appropriate difficulty corresponds to a p-value of approximately .60 (Cronbach & Warrington, 1952; Guilford, 1954; Lord, 1952); items at that difficulty level maximize internal consistency reliability. Thus, maintaining item difficulty near this level for all or most individuals in the group should lead to increased reliability of the measurement tests in comparison to that of the routing test or the linear test, in which items are of median difficulty only for some individuals in the group. Angoff and Huddleston (1958) found this to be the case; their "narrow range" (measurement) tests were more reliable for the groups for which they were intended than were the conventional "broad-range" tests. However, the routing process should also create subgroups of individuals more homogeneous in ability. Because lower ability variance will decrease internal consistency reliability estimates, the effects of more appropriate item difficulties may be counteracted.

Thus, in comparing the internal consistency reliability of the measurement tests to that of the linear and routing tests, it was important, first, to evaluate the extent to which routing led to more optimal measurement test item difficulties; this was done by determining whether item difficulties in the measurement tests changed in the direction of p=.60 from their values as determined from the norming studies. Second, the extent of sub-group homogeneity was evaluated by examining the score variability within each measurement test.

Lord's (1971d) theoretical demonstration that the precision of measurement of two-stage tests was nearly constant over the whole ability range implies fewer random factors in the ordering of individuals in two-stage tests than in conventional tests. In conventional tests, which are most precise at average ability levels, scores of individuals near the extremes of ability will be highly affected by random errors, and the ordering of such individuals will be determined in large part by random factors. Because of the more nearly constant precision of two-stage tests, the scores for individuals at all levels of the ability distribution are more likely to be based largely on underlying ability rather than on random factors; two-stage tests should thus yield higher test-retest stability coefficients than conventional tests. One complicating factor, however, involves differential memory effects. A subject re-tested on the conventional test will repeat the same set of items. A subject retested on the two-stage test will take the same set of 10 routing items but may take an entirely different set of 30 measurement test items if he is routed differently the second time. In comparing the stability, then, of two-stage and conventional tests, it was necessary to account for the differential effects of memory.

Some studies of two-stage testing procedures (e.g., Cleary _et al_., 1968a,b; Linn _et al_., 1969) have evaluated their results in terms of the accuracy with which two-stage test scores estimated scores on a conventional test. The focus of adaptive testing, however, should be on improving the measurement characteristics of scores derived from adaptive tests rather than on estimating conventional test scores. If it is true that two-stage tests yield more precise measurement at the extremes of the distribution than do conventional tests, the ordering of individuals in the tails of the two score distributions should be different. Thus a relatively low correlation with scores derived from a linear test would provide evidence that the two-stage test was ordering individuals _differently_ but would not indicate which ordering had the higher relationship to the trait being measured. Direct evidence pertaining to the latter issue must, of course, come from the examination of each test's relationship to independent ability criteria. Indirect evidence may eventually be derived from determining whether the intercorrelations of a number of adaptive tests, all of which would be constructed to achieve more nearly constant precision throughout the ability range, were uniformly higher than the correlation of each with a conventional test. Analyses pertaining more directly to this issue will be reported in later studies in this series.

## Test Development

### Item Pool

The item pool used to construct the adaptive and conventional tests of verbal ability consisted of 5-alternative multiple-choice vocabulary items. The items were normed on a large group of college students, and item statistics of difficulty (proportion correct) and discrimination (biseral correlation with total score) were obtained. Using a biserial correlation of at least .30 as a selection criterion, 369 items were available for use in constructing the three tests to be administered in the first study. Table 1 describes the available item pool as a cross-classification of levels of item difficulty and biserial correlation coefficient and shows the number of items available in each cell of the cross-tabulation. It may be noted that the pool consisted of considerably more very easy than very difficult items, and that the more highly discriminating items occurred at the easier levels of difficulty.

Two-stage Test

The two-stage test was composed of a 10-item routing test and four 30-item measurement tests. Testees were assigned to one of the four measurement tests on the basis of their scores on the routing test.

Items for each subtest were selected to approximate the characteristics of the theoretical items used by Lord (1971d) in his study of two-stage testing procedures. In describing the characteristics of the theoretical items, Lord used parameters based on assumptions of the normal ogive model in item characteristic curve theory (Lord and Novick, 1968). The characteristics of the real item pool used in this study were specified in terms of the traditional item parameters of classical test theory (i.e., proportion correct as an index of item difficulty, and item-total score correlation as an index of item discriminating power). The normal ogive item parameter values suggested by Lord were used to select the levels of item difficulty and discrmination of the measurement tests. The routing test item difficulties and discriminations were selected by other criteria. Following the selection of the routing and measurement test items, their difficulty and discrimination values were converted to the normal ogive parameters for use in the scoring equation.

Using Lord's notation, normal ogive parameter "a" represents item discriminating power and is related to the biserial correlation between item response and latent ability. Since latent ability estimates were not available for item norming, normal ogive item parameter estimates used in this study were computed using total norming test score as an estimate of latent ability. Although Lord assumed equally discriminating items in his theoretical two-stage tests, he admits it is rarely possible to construct real tests with equally discriminating items. In this study,

Table 1

Number of Vocabulary Items by Item-Test Biserial Correlation and Item Difficulty

| Biserial Item-Total Correlation $r_{it}$ | Item Difficulty (Proportion Correct) | | | | | | | | | | No. of items at each level of $r_{it}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0-<br>.099 | .100-<br>.199 | .200-<br>.299 | .300-<br>.399 | .400-<br>.499 | .500-<br>.599 | .600-<br>.699 | .700-<br>.799 | .800-<br>.899 | .900-<br>.999 | |
| 1.0 | | | | | | | | | | 4 | 4 |
| .90-.99 | | | | | | | | | | 3 | 3 |
| .80-.89 | | | | | | | | | 2 | 8 | 10 |
| .70-.79 | | | | 1 | 1 | 3 | | 1 | 6 | 11 | 23 |
| .60-.69 | 1 | 1 | | 2 | 9 | 5 | 8 | 6 | 10 | 9 | 51 |
| .50-.59 | | 3 | 7 | 7 | 6 | 11 | 18 | 8 | 7 | 15 | 82 |
| .40-.49 | 1 | 6 | 12 | 13 | 8 | 11 | 12 | 14 | 8 | 10 | 95 |
| .30-.39 | 3 | 10 | 15 | 21 | 6 | 17 | 6 | 8 | 6 | 9 | 101 |
| No. of items at each level of p | 5 | 20 | 34 | 44 | 30 | 47 | 44 | 37 | 39 | 69 | 369 |

items were selected whose discriminations clustered as closely as possible around the desired values.

Item parameter "b" represents item difficulty and is essentially a normal distribution transformation of 1-p, although its exact value is dependent on the value of "a". This conversion makes item difficulty more easily interpretable, since positive values correspond to more difficult items and negative values to less difficult items. Lord's two-stage procedures used peaked routing and measurement tests, i.e., all routing items, and all items composing a particular measurement test, had a constant "b" value. Using real items, it was not possible to construct perfectly peaked subtests; rather, desired values of "b" were selected for the measurement tests, and the items were selected to distribute closely around the desired values.

Routing test. The 10 routing items were selected to have a mean item-total score biserial correlation of approximately .57. This value was selected to be somewhat higher than that chosen for the measurement tests in order to improve the assignment of testees to measurement tests.

The difficulty level of the routing items was selected to fall at the median ability level of the group taking into account the probability of chance success on an item as a result of random guessing (Lord's parameter "c"). Lord (1953, 1970) found that optimal measurement could be achieved at a difficulty level somewhat easier than the value of (1+c)/2. Since the items used in this study had 5 alternative responses, "c" was equal to .2, and (1+c)/2 was equal to .60. The mean difficulty level of the routing items was set at .62, slightly easier than p=.60. Thus, ten items with p-values distributed closely around .62 and biserial coefficients as close as possible to .57 were selected for the routing test out of the 369 items available.

The first row of Table 2 summarizes the characteristics of the routing items. The mean, standard deviation, minimum, and maximum values of the traditional item parameters are presented. The mean "a" and "b" values were calculated for use in the scoring equation and are presented after their corresponding traditional item parameter values. It may be noted that the mean biserial correlation (.57) is very close to that desired, but the standard deviation (.07) and range of these values (.43 to .71) show that the items were not equi-discriminating. Similarly, the mean item difficulty fell at the desired point (p=.62), but the 10 items, varying from p=.57 to p=.68, did not form a perfectly peaked test. Item difficulties were normally distributed, with a slight

Table 2

Summary of item characteristics (norming values)
for two-stage and linear tests

| | | Item difficulty | | | | | Item discrimination | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | No. | proportion correct(p) | | | | "b" | biserial correlation | | | | "a" |
| Test | items | Mean | S.D. | high | low | Mean | Mean | S.D. | high | low | Mean |
| Routing | 10 | .62 | .04 | .68 | .57 | -.57 | .57 | .07 | .71 | .43 | .70 |
| Measurement | | | | | | | | | | | |
| 1 | 30 | .24 | .08 | .35 | .09 | 1.75 | .42 | .08 | .67 | .32 | .47 |
| 2 | 30 | .46 | .08 | .58 | .30 | .22 | .44 | .13 | .73 | .31 | .49 |
| 3 | 30 | .73 | .04 | .80 | .63 | -1.34 | .46 | .08 | .61 | .30 | .52 |
| 4 | 30 | .89 | .05 | .96 | .81 | -2.49 | .51 | .12 | .78 | .33 | .59 |
| Linear | 40 | .56 | .08 | .66 | .41 | -.28 | .47 | .06 | .54 | .32 | .54 |

tendency toward flatness rather than peakedness. Appendix Table A-1 shows the characteristics (p-value and biserial coefficient) of each of the 10 routing items.

To make assignments to measurement tests, score ranges on the routing test of 0 through 3, 4 and 5, 6 and 7, and 8 through 10 were used respectively to assign testees to each of four measurement tests. The lowest score range was the widest since it was expected to include many "chance" scores.

Measurement tests. In selecting the measurement test items, a mean item biserial coefficient of .45 was desired. This value corresponds to an "a" of approximately .50, which is the value of item discriminatory power used by Lord in his theoretical studies of adaptive testing (Lord, 1970, 1971a,c,d).

In choosing the difficulty levels of the measurement tests, Lord calculated a value equal to $a(b_2 - b)$, where $b_2$ is the difficulty of a particular measurement test and b is the routing test difficulty. These values were distributed relatively symmetrically around zero and ranged from -1.5 to +1.5 when six measurement tests were available. Because four measurement tests were used in this study, values of +1.0, +.40, -.40, and -1.0 were selected for $a(b_2 - b)$. The corresponding mean item difficulties of the four measurement tests were p=.26, p=.46, p=.73, and p=.88. Thus, in constructing the most difficult measurement test, the 30 items having "p" values closest to .26 and biserial coefficients distributed around .45 were selected; a similar procedure was followed in constructing the other three measurement tests.

The resulting characteristics of the four measurement tests are summarized in Table 2. It may be noted that the mean item difficulties of tests 1 and 4 were slightly different from the desired values; this was due to the necessity of taking item discrimination as well as item difficulty into account. However, the resulting values of $a(b_2 - b)$, which were +1.09, +.39, -.40, and -1.13, were good approximations to the values specified beforehand. As with the routing test, item difficulties of each of the measurement tests were normally distributed around the mean value. Also, the mean biserial correlations for the two most difficult measurement tests were lower than those for the two easier tests. This was due to the relative scarcity of difficult items having high biserial coefficients as was indicated in Table 1. And while the mean biserial levels were relatively close to the .45 value desired, the standard deviation and range of these values show that it was not possible to construct equi-discriminating tests using the available item pool within the limitations of the research design (i.e., the construction of several non-overlapping tests). Appendix Tables A-2 through A-5 give the characteristics of each of the 30 items in each measurement test in terms of p-values and biserial coefficients.

Thus, the two-stage test consisted of a normally distributed routing test whose mean difficulty fell at approximately the median ability level of the group (under the assumptions of random guessing), from which testees were routed or assigned to one of four normally distributed measurement tests whose means were located at points on the ability continuum distributed around the median ability level of the total group.

<u>Scoring</u>. The method used to score the two-stage test was derived from Lord's (1971d) theoretical work. It consisted of obtaining the maximum likelihood estimates of ability from the routing test ($\hat{\theta}_1$ , where $\theta$ indicates position on the latent ability continuum) and the measurement test ($\hat{\theta}_2$). After these two estimates were obtained, they were weighted and then averaged to obtain a composite ability estimate, $\hat{\theta}$. In this study, the estimates of $\theta$ derived from the routing and measurement tests were determined by the following formula:

$$\hat{\theta} = \frac{1}{\bar{a}} \Phi^{-1} \frac{(x/m)-c}{1-c} + \bar{b}$$

In this formula, $\bar{a}$ represents the mean discrimination value of the subtest items, x is the number correct, m is the total number of items administered in that subtest (either 10 or 30), c is the chance-score level (always .2), and $\bar{b}$ represents the mean difficulty of the items in that subtest. Whenever x=m (perfect score) or x=cm (chance score), $\hat{\theta}$ cannot be determined. Therefore, when x was equal to m, it was replaced by x=-.5, and when x was less than or equal to cm, it was replaced by x=cm + .5.

Lord (1971d) admits that there is no uniquely good way to weight the subtest $\hat{\theta}$'s. He computed variance weights, but a preliminary examination of the results of applying his weighting formula to the two-stage data from this study showed some non-monotonicity in the relationship between the number right obtained on the measurement test and the total test $\hat{\theta}$ for people who obtained the same routing score. Therefore, rather than using the variance weights, each subtest $\hat{\theta}$ was weighted according to the number of items on which it was based; the resulting total score estimates were then strictly monotonically related to the actual number correct on the measurement test, given the same routing score. The ability estimate used in this study, then, was defined by the following equation:

$$\hat{\theta} = \frac{(\hat{\theta}_1 \cdot 10) + (\hat{\theta}_2 \cdot 30)}{40} = \frac{\hat{\theta}_1 + 3\hat{\theta}_2}{4}$$

Scores determined in this way have values similar to standard or "z" scores (Lord & Novick, 1968), i.e., most will fall between $\pm 3$, and the meaning of a $\hat{\theta}$ of +1 corresponds to that of a standard or "z" score of +1.

In the following sections, references to "two-stage" scores will always refer to $\hat{\theta}$; scores reported for the routing and measurement tests, on the other hand, will always refer to the number correct on the particular subtest in question.

Conventional linear test. Lord (1971d) compared his 60-item two-stage tests with a 60-item peaked linear test having equi-discriminating items (biserial correlations with the underlying trait of about .45). The linear test used for comparative purposes in this study had 40 items so that its length would equal that of the two-stage test. Items were selected from the pool shown in Table 1 that had difficulties closest to p=.55 and item-total score biserial correlation coefficients closest to .45. The mean, standard deviation, minimum value, and maximum value of the linear test item difficulties and biserial coefficients are shown in Table 2. Again, the mean values of the normal ogive parameters are presented for comparative purposes. As was true for the routing and measurement tests, the linear test was neither equi-discriminating nor perfectly peaked. The linear test did have a smaller range of item biserial values (.32 to .54) than did the two-stage subtests, and the range of item difficulties (.41 to .66), while large for a peaked test, was small in relation to the range covered by all of the four measurement tests. The distribution of linear test item difficulties, like that of the two-stage subtests, was normal. Appendix Table B-1 presents the p-value and biserial coefficients for each of the 40 items in the linear test.

An individual's score on the linear test was simply the number of correct responses given to the 40 items; thus scores could potentially vary from 0 and 40.

## Administration and Subjects

The tests were administered to undergraduate students taking the introductory psychology and basic psychological statistics courses at the University of Minnesota. The students were tested at individual cathode-ray-terminals (CRTs) connected by acoustical couplers to a time-shared computer. The CRTs were located in quiet rooms, and there was a maximum of 3 students in each room at one time. An administrator was present at all times to help students with the terminal equipment and to ensure that no consultation took place among testees. A set of instructional screens preceded the beginning of testing on all of the initial tests, and the students were given the opportunity to review the instructional screens before taking the retest. Few students had difficulty operating the terminals after completing the instructions; CRT test administration thus seems quite appropriate for college students.

On the first testing, 214 students completed the two-stage test and of these 112 also took the linear test (the remainder completed a flexilevel test). The students were retested after a mean interval of 39 days (about $5\frac{1}{2}$ weeks), with a standard deviation of 11 days and a range from 14 to 62 days. Of the 214 students who completed a two-stage test on first testing, 178 were retested, and of these 85 also completed the linear test a second time (the remainder completed another adaptive test on retest).

## Analysis of Data

The data to be analyzed consisted of 2 two-stage test scores, one from the initial test (time 1) and one from the retest (time 2), for each individual. For about half of the group there were also 2 scores (test and retest) from the linear test. The time 1 data was divided into 2 groups, one consisting of those subjects who had taken the two-stage test first and the linear test second (order 1) and the other consisting of those subjects for whom the order was reversed (order 2). To analyze the effect of order of administration, mean scores from order 1 and order 2 for the two-stage test and the linear test were compared using a t-test of the significance of mean differences. Table 3 presents the score means and standard deviations derived from order 1 and order 2 and the value of t and its associated probability for each comparison. Since there were no significant differences

Table 3

Means and standard deviations of test scores for analysis of order effects, and t-tests of the significance of mean differences

| | Order 1 | | | Order 2 | | | Test of Significance | | |
|---|---|---|---|---|---|---|---|---|---|
| | Two-stage--Linear | | | Linear--Two-stage | | | t | degrees of | |
| Test | N | Mean | S.D. | N | Mean | S.D. | value | freedom | p |
| Two-stage | 115 | -.27 | 1.26 | 99 | -.15 | 1.47 | -.66 | 212 | .51 |
| Linear | 59 | 23.80 | 8.42 | 53 | 24.62 | 8.19 | -.53 | 110 | .60 |

between means for either the two-stage or linear tests, order of administration was concluded to be an unimportant variable, and all subsequent analyses were done with data from the two order groups combined.

## Characteristics of Score Distributions

Analyses of the characteristics of the score distributions were done separately for initial test data and for retest data. The score means and standard deviations were calculated for each distribution, but because the scores were expressed in different terms (i.e., number correct for the linear test versus position on a latent ability continuum for the two-stage), the scores and their means and standard deviations were not directly comparable.[1] Thus, in order to compare the variability of the score distributions, an index of relative variability was computed. This index indicates the extent to which the potential score range is effectively utilized and was computed by dividing the standard deviation of each score distribution by its total potential score range. The score range for the linear test was 40, and that for the two-stage test was 6 (±3 standard deviations on the latent ability continuum).

To determine the nature of the score distributions, measures of skewness and kurtosis were obtained and tested for significant departures from normality (McNemar, 1969, pp. 25-28 and 87-88).

## Reliability

*Internal consistency*. Internal consistency reliability for the linear test and for each subtest (i.e., routing test and the four measurement tests) of the two-stage test was estimated by the Hoyt (1941) method. However, since the reliabilities of the linear test, the routing test, and the measurement tests were based on different numbers of items, they were not directly comparable. Thus, the Spearman-Brown prophecy formula was used to project the reliabilities of the two-stage subtests to what they would be had they been based on 40 items (the length of the linear test) rather than 10 items (routing) or 30 items (measurement).

To determine whether or not the measurement test item difficulties were appropriate for maximizing internal consistency, the mean difficulty of the items in each measurement test for that group of subjects who had taken it was calculated. For further comparisons of the item statistics

[1] The linear test scores could also have been expressed in terms of $\theta$, or position on the latent ability continuum. However, since most conventional tests are scored using "number correct", that scoring method was used in this study to maintain practical relevance of the results.

as derived from the norming and the actual test administration, the means and standard deviations of the discriminations (biserial correlation with total score) of the measurement test items were calculated. The item difficulty and discrimination statistics were also calculated for the linear and routing tests. The total score used in these calculations was the number correct score on the linear test, and the number correct on the two-stage subtest rather than $\hat{\theta}$. The item statistics for the linear and routing tests were based, of course, on the total group of testees, whereas those for the measurement tests were based only on that more homogeneous group of testees who had completed each measurement test.

To determine the extent to which the routing process had led to a restriction of range, or greater homogeneity of ability, within each measurement test subgroup, the means and standard deviations of the number correct scores on each measurement test, and also on the linear and routing tests, were calculated. To facilitate comparison of the standard deviations, which were based on tests of 10, 30, or 40 items, each standard deviation was divided by its total potential score range (the number of items in the test) to obtain the index of the extent to which the potential score range was used.

Stability. A series of analyses of test-retest stability were done. First, Pearson product-moment correlation coefficients were calculated for the test-retest score distributions of each test. Eta coefficients and the significance of curvilinear relationships between the test and retest scores were also calculated. Second, to examine the effect of interval length on test-retest stability, the total group was divided into three subgroups according to the length of interval between test and retest. The three groups were short interval (14-30 days), moderate interval (31-46 days), and long interval (47-62 days); product-moment correlation coefficients were then calculated for the test-retest scores of the individuals in each subgroup.

Third, in order to analyze the effect of memory of the items on test-retest stability, two-stage stability coefficients were calculated using only those individuals who were routed into the same measurement test on both testings. These individuals thus took the same 40 items on test and retest, therefore making the effects of memory comparable to that of the linear test, on which all subjects repeated the same 40 items.

## Additional Analyses

To analyze the relationship between the two-stage and linear test scores, product-moment correlations and eta coefficients for each total score distribution regressed on the other one were computed. Tests of curvilinearity were made to determine if there were non-linear relationships between the two score distributions.

Other analyses concerned certain characteristics of the two-stage test itself. First, the distribution of routing test scores and the number and percentage of individuals assigned to each measurement test were examined in order to evaluate the appropriateness of the difficulty level of the routing test and the score intervals selected for assigning testees to measurement tests. Second, the number and percentage of misclassifications into measurement tests was determined; the criteria selected to identify misclassified individuals were 1) perfect scores (all 30 items correct), indicating that the measurement test was too easy, and 2) chance scores (6 or less correct responses), indicating that the test was too difficult.

# RESULTS

## Comparison of Two-stage and Linear Tests on Psychometric Characteristics

Variability. Table 4 presents the means, standard deviations, and the "proportion of range utilized" index of variability for the two-stage and linear test scores. The data in Table 4 show that the two-stage scores utilized a slightly larger proportion of their potential range than did the linear test scores, on both the original testing and the retest. Further, although the mean scores on both tests increased on the retest, the standard deviations and the proportion of range utilized were the same on original testing and on retest for both the two-stage and linear test scores, thus suggesting consistency in the extent to which scores derived from each test utilized the available score range.

Shape of the score distributions. Table 5 presents data describing the two-stage and linear score distributions. The two-stage distributions, for both test and retest, satisfied the criteria of normality, since neither the indices of skewness nor kurtosis were significantly different from zero. However, there was some tendency toward positive skew and flatness in both distributions of

Table 4

Mean, standard deviation, and "proportion of range utilized" index of variability for two-stage and linear test scores

| | Time 1 | | | | Time 2 | | | |
|---|---|---|---|---|---|---|---|---|
| Test | N | Mean | S.D. | Proportion of range utilized | N | Mean | S.D. | Proportion of range utilized |
| Two-stage | 214 | -.21 | 1.36 | .23 | 178 | -.02 | 1.39 | .23 |
| Linear | 110 | 24.19 | 8.28 | .21 | 85 | 25.67 | 8.32 | .21 |

Note: Proportion of range utilized is calculated by dividing the standard deviation by the potential score range.

Table 5

Indices of skewness and kurtosis and associated standard errors for score distributions of two-stage and linear tests

| | Time 1 | | | | | Time 2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Test | N | Skew | S.E. | Kurtosis | S.E. | N | Skew | S.E. | Kurtosis | S.E. |
| Two-stage | 214 | .27 | .18 | -.09 | .33 | 178 | .28 | .18 | -.52 | .37 |
| Linear | 110 | -.04 | .23 | -1.01* | .46 | 85 | -.24 | .26 | -.93 | .52 |

*significant at p < .02

two-stage scores. The linear test scores, on the other hand, showed some tendency, although not statistically significant, toward negative skew and showed a marked tendency toward flatness on the initial test. The latter result was statistically significant at the .02 level.

## Reliability

Internal consistency. Table 6 presents the Hoyt internal consistency reliability coefficients for the linear test and each two-stage subtest, and the estimated reliability of each subtest had its length been 40 items. It is evident that the linear test and the "40-item" routing test were highly reliable and more reliable than any of the measurement tests. The two intermediate difficulty measurement tests (tests 2 and 3) had especially low reliability coefficients. These findings are contrary to those of Angoff and Huddleston (1958), who found that the measurement ("narrow-range") tests were more reliable than the conventional ("broad-range") test. The results are also contrary to the expectation that higher reliabilities would result from more appropriate item difficulties, i.e., item difficulties close to .60, the median difficulty with chance taken into account, in each measurement test.

Table 7 shows the mean item difficulties for each two-stage subtest and the linear test. The means for the linear test, both time 1 and time 2 (.60 and .64) were very close to .60, and those for the routing test (.68 and .71) although somewhat easier, were still relatively close to .60. On the other hand, with the exception of test 3, the measurement tests were not maximally appropriate for the groups taking them, since their mean item difficulties were not close to p=.60. Measurement test 4 was obviously too easy for those routed to it (p=.78 and .81) while measurement test 1 (p=.43 and .44) was too difficult.

However, in addition to the fact that three of the four measurement tests were not of optimal difficulty, there was evidence for a restriction of range or decreased group heterogeneity and, thus, depressed internal inconsistency reliability coefficients. Table 8 shows the means and standard deviations of the number correct scores for the two-stage subtests and the linear test and the standard deviations as proportions of the number of items (potential range) in each test. As is shown, the proportion of potential range used by the 10-item routing test (.23 on both test and retest) was somewhat greater than that used by the 40-item linear test (.21 both times). But the

Table 6

Internal consistency reliability of routing test, measurement tests,
and linear test, and estimated reliability for
40-item routing and measurement tests

| | Time 1 | | | | Time 2 | | | |
|---|---|---|---|---|---|---|---|---|
| Test | N | Hoyt reliability coefficient | No. of items | Estimated reliability for a 40-item subtest | N | Hoyt reliability coefficient | No. of items | Estimated reliability for a 40-item subtest |
| Routing | 214 | .68 | 10 | .89 | 178 | .69 | 10 | .90 |
| Measurement | | | | | | | | |
| 1 | 91 | .79 | 30 | .83 | 93 | .78 | 30 | .82 |
| 2 | 61 | .66 | 30 | .72 | 41 | .62 | 30 | .69 |
| 3 | 39 | .44 | 30 | .51 | 28 | .50 | 30 | .58 |
| 4 | 23 | .82 | 30 | .86 | 16 | .70 | 30 | .78 |
| Linear | 110 | .89 | 40 | .89 | 85 | .90 | 40 | .90 |

Table 7

Mean and standard deviation of item difficulties (proportion correct) obtained from administration of the two-stage and linear tests

| | Proportion correct | | | | | |
|---|---|---|---|---|---|---|
| | Time 1 | | | Time 2 | | |
| Test | No. items | Mean | S.D. | No. items | Mean | S.D. |
| Routing | 10 | .68 | .12 | 10 | .71 | .09 |
| Measurement | | | | | | |
| 1 | 30 | .43 | .16 | 30 | .44 | .15 |
| 2 | 30 | .51 | .11 | 30 | .47 | .12 |
| 3 | 30 | .64 | .15 | 30 | .69 | .11 |
| 4 | 30 | .78 | .13 | 30 | .81 | .13 |
| Linear | 40 | .60 | .11 | 40 | .64 | .12 |

Table 8

Mean, standard deviation, and standard deviation as proportion of potential range (number of items) for the two-stage subtests and the linear test

| | Time 1 | | | | Time 2 | | | |
|---|---|---|---|---|---|---|---|---|
| Test | N | Mean | S.D. | S.D./No. of items | N | Mean | S.D. | S.D./No. of items |
| Routing | 214 | 6.78 | 2.31 | .23 | 178 | 7.18 | 2.28 | .23 |
| Measurement | | | | | | | | |
| 1 | 91 | 12.98 | 5.28 | .18 | 93 | 13.34 | 5.25 | .18 |
| 2 | 61 | 15.38 | 4.51 | .15 | 41 | 14.10 | 4.28 | .14 |
| 3 | 39 | 19.13 | 3.33 | .11 | 28 | 20.79 | 3.48 | .12 |
| 4 | 23 | 23.39 | 4.81 | .16 | 16 | 24.19 | 3.82 | .13 |
| Linear | 110 | 24.19 | 8.29 | .21 | 85 | 25.67 | 8.32 | .21 |

measurement tests, which had 30 items, used considerably less of the potential range than did either the routing test or the linear test. Measurement test 3 used only half as much of its potential score variability as did the linear and routing tests. Referring back to Table 6, it is interesting to note that the reliability coefficients are very closely related to the proportions of potential range used by each of the tests. For example, measurement test 3 was both the least variable and the least reliable. In general, the rank order of the tests or subtests in terms of internal consistency reliability corresponds to their rank order in terms of score variability. Thus, it would seem that the increased homogeneity of the groups of subjects taking each measurement test, as evidenced by the low score variability, was an important factor in the unreliability of the measurement tests.

The low score variability of the measurement tests in comparison to that of the linear test is in contrast with the comparatively high variability of the total scores on the two-stage test as was shown in Table 4. However, given the fact that the testees were all college undergraduates, a group that can be assumed to have an already restricted range of ability from that in the general population, it is not surprising that dividing this total group into four subgroups even more homogeneous in ability led to reduced score variability. It is likely that the measurement tests would show higher reliability if the two-stage test were administered to a group more representative of the general population in terms of a greater range of ability levels.

Stability. Table 9 gives the test-retest stability correlations for the two-stage and linear tests. The first three sets of columns show the stability correlations as a function of the length of the interval between test and retest; the last two columns show the stability of each test as computed on the total group of subjects.

The length of the interval between test and retest did not have consistent effects on stability. The linear test was most stable in the interval of medium length (r=.91) and least stable in the longest interval (r=.87), whereas the two-stage test was most stable in the shortest interval (.92) and least stable in the medium-length interval (.85). It is interesting, though possibly not significant, to note that the two-stage test was more stable over the longest interval than the linear test. This may have some implications for the relative importance

Table 9

Test-retest stability correlations as a function of interval length, and for total group

| | Retest Interval (in days) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 14-30 | | 31-46 | | 47-62 | | Total group | |
| Test | N | r | N | r | N | r | N | r |
| Linear | 25 | .89 | 28 | .91 | 21 | .87 | 74 | .89 |
| Two-stage | 41 | .92 | 66 | .85 | 47 | .89 | 154 | .88 |

of memory effects in the stability of the two tests, i.e., if memory of the items is important in the stability of a test, the longer the interval, the less effect memory will have and, thus, the lower will be the stability coefficient.

The linear test (r=.89) had a slightly higher total group stability than the two-stage test (r=.88), but the difference was not significant and could easily have been in the opposite direction. Tests for curvilinearity, using the product-moment correlations and eta coefficients, showed that the relationship between the test and retest scores was primarily linear, with no significant curvilinearity.

In addition to the effect of interval length on the obtained test-retest stability coefficient, the other factor considered was the effect on the size of the stability coefficient of memory of the items on the retest. The stability of the linear test, which was r=.89, was based on the correlation between the test and retest scores of subjects who had repeated the same 40 items. The stability of the two-stage scores was, therefore, calculated only for the 97 subjects who were assigned to the same measurement test on both test and retest, thus also repeating the same 40 items. That test-retest stability correlation was .93, higher than both the linear and the total group two-stage stability coefficients. Thus it would appear that the stability of the linear test was based to a larger extent on memory of the items than was that of the two-stage test, suggesting that the latter yields ability estimates which more consistently reproduce the testee's ability over the time interval between testings.

Relationships between Linear and Two-stage Scores

Table 10 presents the linear (product-moment) and eta coefficients describing the relationships between the two-stage and linear score distributions on test and retest. All of the linear and eta coefficients were significant at $p \leq .001$. The only significant degree of curvilinearity was found in the regression of the linear scores on the two-stage scores for the initial test, although there was a tendency toward curvilinearity (p=.12) in the regression of two-stage on linear scores on the retest. Examination of the bivariate scatter plots showed that the curvilinearity was due to a restriction of range in the lower end of the linear score distribution in comparison to the greater utilization of the two-stage score range at the lower ends.

The linear relationship between the two-stage and linear test scores was relatively high on both test and

Table 10

Regression analysis of relationships between two-stage scores and linear scores, and tests for curvilinearity (N=110 Time 1, N=85 Time 2)

| | Time 1 | Time 2 |
|---|---|---|
| Product-moment correlation | .84 | .80 |
| Eta coefficients | | |
| Regression of two-stage scores on linear scores (eta) | .85 | .84 |
| Significance of curvilinearity (p-value) | .74 | .12 |
| Regression of linear scores on two-stage scores (eta) | .88 | .82 |
| Significance of curvilinearity (p-value) | .04 | .90 |

retest (.84 and .80). However, these values also indicate that the proportions of variance accounted for ($r^2$) were only .70 and .64, respectively. The proportions of reliable variance in the linear test, as given by the Hoyt internal consistency reliability coefficients, were .89 and .90; thus, the correlation between the two-stage and linear test scores failed to account for 19% of the reliable variance in the linear test on initial testing, and 26% on retest. It would appear, therefore, that the linear test and the two-stage test are not interchangeable approaches to measuring the same ability.

## Comparison of Norming and Testing Item Statistics

Since this study is the first to report on non-simulated two-stage test administration, it is appropriate to examine the effect of actual two-stage testing on item characteristics. Relevant data from both the two-stage and linear test have been presented earlier in Table 7; additional data are in Tables 11 and 2.

Item difficulties. Table 7 gives the means and standard deviations of item difficulties as obtained from actual administration of the two-stage and linear tests. These values may be contrasted with the values as obtained from the norming studies, which were presented in Table 2.

It may be noted, first of all, that the linear and routing tests, both of which were taken by the total group of subjects, were somewhat easier for the tested group (on first testing) than they had been for the norming sample. On the linear test, average difficulty for the norming group (Table 2) was p=.56, while for the tested group (Table 7) it was p=.60 (time 1). On the routing test the respective average difficulties were p=.62 for the norming group and p=.68 for the tested group. Since both of these differences were statistically significant ($p < .05$), it is possible that the tested group was slightly superior in verbal ability, although both samples were taken from the same population.

However, of more importance in this study was the effect that changes in group composition toward greater homogeneity in ability level, caused by the routing process, would have on the item difficulties of the measurement tests. On all four measurement tests, the testing mean item difficulties changed in the direction of p=.60 from their norming values. The two more difficult measurement tests (1 and 2), with norming means of .24 and .46, were significantly easier ($p < .001$ and $p < .01$) and closer to median difficulty for the groups of testees routed into them ($\overline{p}$=.43 and .51

Table 11

Mean and standard deviation of item discrimination values (biserial correlation with total number correct) from administration of the two-stage and linear tests

| | | Time 1 | | Time 2 | |
|---|---|---|---|---|---|
| | No. | Biserial coefficient | | Biserial coefficient | |
| Test | items | Mean | S.D. | Mean | S.D. |
| Routing | 10 | .67 | .10 | .69 | .11 |
| Measurement | | | | | |
| 1 | 30 | .49 | .14 | .46 | .16 |
| 2 | 30 | .39 | .19 | .37 | .18 |
| 3 | 30 | .31 | .19 | .37 | .25 |
| 4 | 30 | .60 | .32 | .44 | .42 |
| Linear | 40 | .56 | .15 | .58 | .16 |

respectively). Similarly, the two less difficult tests (3 and 4), with norming values of .73 and .89, were significantly more difficult ($p < .05$ and $p < .001$) and closer to median difficulty for the subjects taking them ($\overline{p}$=.64 and .78 respectively). These findings suggest that each measurement test was more appropriate to the ability level of that subgroup taking it than it would be for the total group of subjects.

Tables 2 and 7 also show that the testing values of the standard deviations of the item difficulties were uniformly larger than the norming values. This finding implies that groups of items which show very similar characteristics when normed on one group of subjects may show more divergent characteristics when administered to groups differing from the norming sample in composition and range of ability levels.

Item discriminations. Table 11 presents the means and standard deviations of item discrimination values (biserial correlation with number correct) as obtained from the administration of the tests. A comparison of these values with the norming values as presented in Table 2 shows that the testing mean item discrimination values for the linear and routing test were higher than the corresponding norming values; the mean biserials of the linear test items were .47 from the norming studies but .56 and .58 from the test and retest, and the routing test increased from a mean discrimination of .57 in norming to .67 and .69. In contrast, the only measurement test to show higher item discrimination values on both test and retest was test 1, the most difficult test, whose means were .42 in norming but .49 and .46 on test and retest. The items in tests 2 and 3 were less discriminating in testing than they had been in norming, and those in test 4 were more discriminating on the first test but less discriminating on the retest. Further, the standard deviations of the item discrimination values were again larger in testing than they had been in norming. The items in test 4 especially showed much greater variability in their discriminating power.

The substantial changes that were found in both the level and variability of item discriminating power were probably a factor in the rather poor internal consistency reliability of the measurement tests and suggest that item statistics derived from norming samples composed of one range of ability levels may be inappropriate when applied to a group composed of a different range of ability levels.

## Additional Characteristics of the Two-stage Test

The results thus far have suggested certain problems with the two-stage test. Three of the four measurement tests were not of optimal difficulty for the groups of subjects taking them, and the item discrimination values of the measurement tests tended to be both lower and more variable in actual two-stage testing than they had been in norming. Thus, the two-stage test was further examined to evaluate the degree to which it met its major objective. That is, the two-stage test was analyzed to determine whether the "routing" test assigned members of a group of individuals varying rather widely in ability to longer "measurement" tests such that each measurement test was essentially "peaked" at the mean ability of a far more homogeneous group of subjects and was thus more appropriate to their level of ability than would be a test designed to measure the full range of ability within the larger group.

In first examining the characteristics of the 10-item routing test, it was found that the mean number correct was 6.78 on the first test and 7.18 on the retest (see Table 8). These high mean scores were close to expectation because the test was constructed to be somewhat easier than the median ability with chance success accounted for (p=.60). However, on both test and retest, the distribution of routing test scores showed a significant degree of negative skew, indicating a predominance of high scores (7 to 10 correct).

The high and significantly skewed routing scores, coupled with the score intervals selected for assignment to measurement tests (0-3, 4-5, 6-7, and 8-10), meant that a majority of the testees were assigned to the two most difficult measurement tests (tests 1 and 2). Table 12 summarizes data on the number and percentage of the total group assigned to each measurement test and the mean and standard deviation of the number correct scores obtained by each of these subgroups.

The data in Table 12 show several deficiencies of the two-stage test used in this study. First, the imbalance in the numbers of testees taking the individual measurement tests is obvious and consistent; roughly half of the total group took the most difficult test on both test and retest, whereas only about one-tenth of the group took the easiest test. Although the percentages taking each test time 1 and time 2 are fairly comparable, there was a tendency for the imbalance to be even more pronounced on the retest.

Table 12

Number and percentage of total group assigned to each measurement test and mean and standard deviation of number correct (of 30 possible) for each test

| | | Time 1 | | | | Time 2 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Number correct | | | | Number correct | |
| Measurement test | Score range on routing test | N | % | Mean | S.D. | N | % | Mean | S.D. |
| 1 | 8-10 | 91 | 42.5 | 12.98 | 5.28 | 93 | 52.2 | 13.34 | 5.25 |
| 2 | 6-7 | 61 | 28.5 | 15.38 | 4.51 | 41 | 23.0 | 14.10 | 4.28 |
| 3 | 4-5 | 39 | 18.2 | 19.13 | 3.33 | 28 | 15.7 | 20.79 | 3.48 |
| 4 | 0-3 | 23 | 10.7 | 23.39 | 4.81 | 16 | 9.0 | 24.19 | 3.82 |

Second, as was pointed out in the section on reliability, the tests were not of optimal difficulty for those groups of individuals taking them. The most appropriate mean item difficulty would be around p=.60, meaning that the desired mean number correct on each measurement test would be about 18. As Table 12 shows, however, the two most difficult tests were too difficult (mean total scores of 12.98 and 15.38 respectively) for the average subject taking them, and the two easier tests were too easy (means of 19.13 and 23.39 respectively). These results and the findings of the rather low number-correct score variability of the measurement tests, as shown in Table 8 and discussed in the reliability section, suggest that the total group was more homogeneous in ability than expected. If the cutting scores for assignment to measurement tests had been set higher, e.g., 0-4, 5-6, 7-8, and 9-10, the two most difficult measurement tests would probably have been more appropriate, but the placement of higher ability subjects into the easier tests would have made these two tests even easier, and thus more inappropriate for many of the individuals assigned to them, than they were using the score intervals selected for this two-stage test.

Misclassification. A different approach to the evaluation of the appropriateness of assignment to measurement tests was to identify the extent to which particular individuals were classified into inappropriate tests. Defining misclassified individuals as those who obtained perfect scores (e.g., all 30 items correct), indicating that the test was too easy, or scores at or below chance (i.e., scores of 6 or less correct), indicating that the test was too difficult, there were 9 or 4.2% misclassifications on the first test and 9 or 5.0% on the retest. All 18 misclassifications were the result of scores at or below chance on the most difficult measurement test, thus providing additional evidence that this test was too difficult for many individuals routed to it. However, the 4 to 5% misclassification rate obtained here was a considerable improvement over the 20% rates obtained in the studies of Angoff and Huddleston (1958) and Cleary *et al*. (1968a,b), although this may be due in part to different criteria of misclassification. Thus, although the measurement tests were not optimal for the *groups* taking them, few *individuals* took a test which was highly inappropriate.

## CONCLUSIONS AND IMPLICATIONS

Considering that the two-stage test used in this study had some deficiencies, the findings of the study were generally favorable to the continued exploration of two-stage testing procedures. The two-stage test, scored using a variation of the method used in Lord's (1971d) theoretical study, yielded scores which were normally distributed and utilized a consistently higher proportion of the available score range than did the linear test. In other empirical studies of adaptive testing where the distribution of scores has been examined, a tendency toward badly skewed scores with definite bunching at the high end of the distribution has been found (Bayroff & Seeley, 1967; Bayroff, Thomas & Anderson, 1960; Seeley, Morton, & Anderson, 1962). Thus, the two-stage test constructed for this study yielded a better distribution of scores than has been found in most empirical studies of adaptive testing to date. The significantly flat distribution of linear test scores may have been a function of deviations from peakedness in its construction; a more peaked test might have yielded a more normal distribution of scores.

The findings regarding the reliability of the two-stage test were less clear. In terms of test-retest stability, the two-stage test scores were quite reliable (r=.88) over a mean interval of 5.5 weeks, essentially as stable as the linear test scores (r=.89). However, when the effect of memory of the items was equated for the two testing strategies, the two-stage scores were the more stable (r=.93). Thus, the two-stage test yielded 7.3% more stable variance than did the linear test of the same number of items and with the same potential for memory effects.

The relatively poor internal consistency reliability of the measurement tests, as compared to the high reliabilities of the routing test and the conventional linear test, was a finding in contrast to those of Angoff and Huddleston (1958) and was probably due to a combination of factors. First, the routing process created subgroups of individuals who were very homogeneous in ability. This was not an unexpected finding, especially given the relative homogeneity of ability in a college student population in comparison to that in a more general population. Further, even though increasing subgroup homogeneity decreases internal consistency, the purpose of the two-stage test is to do precisely that; by initially classifying a group of subjects as to ability, as the routing test does, it is possible to measure them using the most appropriate peaked measurement test. The best two-stage testing procedure would be one containing an infinite number of measurement tests, such that there would be a peaked test perfectly suited to each individual's ability. In this hypothetical mode of testing, there would be complete homogeneity of ability within subgroups since each measurement test would be taken only by individuals with exactly equal ability. Thus, it is perhaps unrealistic to expect high internal consistency reliability from tests which function in this way.

In addition to the extreme subgroup homogeneity, the item difficulties of the measurement tests were not optimal for high reliability, and many of the items which had been highly discriminating in the norming studies were much less discriminating when administered to more homogeneous samples from the total group, thus reducing the internal consistency. Both of these inadequacies can be traced to the inappropriateness of traditional methods of determining item parameters for items to be used in adaptive testing. Only after administering a two-stage test to a defined group of individuals is it possible to determine how difficult and how discriminating the items will be for each subgroup of individuals formed; thus, selecting items for two-stage tests using traditional item parameters can at present be only an approximate procedure. Perhaps the construction of future two-stage tests should use item parameters derived from heterogeneous samples for selection of the routing test items but item parameters derived from more homogeneous subgroups of the total norming sample for the selection of items for each of the measurement tests. Alternatively, item parameters estimated using the techniques of modern test theory (e.g., Lord & Novick, 1968) might be appropriate if it can be shown that these parameters are independent of the range and level of ability in the groups on which they are determined.

The selection of score intervals for assignment to measurement tests is also a matter that needs further study. In this study, the score intervals selected were somewhat inappropriate, leading to an uneven distribution of testees among measurement tests. Although the measurement tests were more appropriate in difficulty for the groups taking them than a test peaked at the median total-group difficulty would be, they were still either somewhat too easy or somewhat too difficult for the groups taking them. However, few individuals were misclassified under the criteria used; the 5% rate of misclassification was a large improvement over the 20% rates of Angoff and Huddleston's (1958) and Cleary et al.'s (1968a,b; Linn et al., 1969) two-stage tests.

The relationship between the linear and two-stage test scores was relatively high (.84 and .80) and primarily linear. The nonlinearity that was found in the regression of the linear scores on the two-stage scores on the first test seemed to be due to restriction in the lower score ranges of the linear test in comparison to the lack of range restriction in the two-stage scores. However, further analyses showed that the relationship between the two tests left about 20% of the reliable variance in the linear test scores and an unknown amount of reliable variance in the two-stage test scores unaccounted for.

A conventional linear test, however, should not be taken as a standard against which new methods of testing must be evaluated. Although a peaked conventional test provides probably the most accurate measurement for individuals whose ability level is near the group mean or the difficulty level at which the test is peaked, its accuracy becomes increasingly less as an individual's ability level deviates from the mean (Lord, 1970, 1971a,c,d). Adaptive tests, on the other hand, provide almost constant accuracy throughout the range of ability (Lord, 1970, 1971a,c,d). Thus, the relationship between the two-stage and linear tests can become meaningful only in the comparative context of indices of relationship between other adaptive strategies and the two-stage test, and indices of the extent to which the two-stage test and the linear test are found to predict a variety of relevant external criteria. Previous studies of two-stage and other adaptive testing strategies have found the adaptive tests to have higher relationships with external criteria than conventional tests of equivalent length (Angoff and Huddleston, 1958; Linn _et al_., 1969; Waters, 1964, 1970; Waters & Bayroff, 1971; see Weiss & Betz, 1973). No studies to date have examined the relationships between two or more adaptive tests. Thus, the validation of two-stage testing procedures depends on additional research in this area.

For further study of two-stage testing procedures, it should be possible to use the information gained in this study to select more optimal score intervals for assignment to measurement tests, to select more appropriate measurement test item difficulties, and to improve the internal consistency reliability by selecting items shown to be highly discriminating for particular subgroups as well as for the total group. A method of selecting the routing test score intervals that would probably be superior to rational or trial-and-error selection would be to compute each individual's latent ability estimate from the routing test ($\hat{\theta}_1$, as described in the scoring section) and to assign him to that measurement test whose mean difficulty in normal ogive parameter terms ("b" values) is closest to the estimate of his/her ability derived from the routing test.

However, the most obvious deficiency of two-stage testing procedures in general is that individuals may be routed to highly inappropriate measurement tests. A low ability individual may guess enough routing items correctly to place him in a measurement test that is too difficult. A higher ability individual confronted with a set of routing items that he is unable to answer correctly as a result of specific gaps in his knowledge or anxiety at the early stages of testing would be routed to a measurement test that is too easy.

One approach to this problem, of course, would be to lengthen the routing test. This approach, however, would undermine one advantage of two-stage testing, i.e., to arrive at an initial estimate of each individual's ability as quickly and efficiently as possible so that a larger set of items relevant to his/her ability may be administered. A more desirable approach would seem to be to include a recovery routine in the computer program controlling test administration. This routine would detect individuals who had apparently been misclassified after only a few measurement test items had been administered; for example, a chance score or a near-perfect score after 10 measurement test items had been administered would cause the individual to be re-routed into the next easier or next more difficult measurement test. The process could be repeated if following re-routing the individual was still wrongly classified. This procedure would mean that individuals would complete different total numbers of items depending on the ease or difficulty of correctly classifying them; thus, the number as well as the difficulty level of the items administered would be adapted to each individual.

Much empirical research remains to be done on two-stage testing procedures; if the information gained from previous empirical studies and the possibilities for improvements suggested by these studies can be fully utilized in subsequent research, it is likely that two-stage testing procedures will become valuable and practical alternatives to traditional testing procedures.

## References

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

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

Bayroff, A. G. Feasibility of a programmed testing machine. U. S. Army Personnel Research Office, Research Study 64-3, 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.

Cleary, T. A., Linn, R. L., & Rock, D. A. An exploratory study of programmed tests. Educational and Psychological Measurement, 1968, 28, 345-360. (a)

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

Cronbach, L. J. & Gleser, G. C. Psychological tests and personnel decisions. (2nd Ed.) Urbana: University of Illinois Press, 1965.

Cronbach, L. J. & Warrington, W. G. Efficiency of multiple-choice tests as a function of spread of item difficulties. Psychometrika, 1952, 17, 127-147.

Evans, R. N. A suggested use of sequential analysis in performance acceptance testing. Urbana: College of Education, University of Illinois, mimeo, 1953.

Guilford, J. P. Psychometric methods. New York: McGraw-Hill, 1954.

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

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

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

Lord, F. M. The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika*, 1952, *17*, 181-194.

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

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

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

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

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

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

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

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

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

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

Statistical Research Group, Columbia University. Sequential analysis of statistical data, applications. New York: Columbia University Press, 1945.

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

Wald, A. Sequential analysis. New York: Wiley, 1947.

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

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

Weiss, D. J. Individualized assessment of differential abilities. Paper presented at the 77th Annual Convention of the American Psychological Association, Division 5, September, 1969.

Weiss, D. J. Strategies of computerized ability testing. Research Report 73-x, Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis. (in preparation)

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

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

Wood, R. Response-contingent testing. Review of Educational Research, 1973 (in press).

# Appendix A

## Item Specifications for Two-stage Test

Table A-1

Item difficulty and discrimination indices for the Routing Test

| Item No. | Difficulty<br>(p) | Discrimination<br>($r_b$) |
|---|---|---|
| 1 | .568 | .708 |
| 2 | .566 | .653 |
| 3 | .589 | .563 |
| 4 | .635 | .608 |
| 5 | .626 | .552 |
| 6 | .622 | .552 |
| 7 | .675 | .566 |
| 8 | .674 | .554 |
| 9 | .677 | .547 |
| 10 | .598 | .430 |

Table A-2

Item difficulty and discrimination indices for Measurement Test 1

| Item No. | Difficulty (p) | Discrimination ($r_b$) |
|---|---|---|
| 1 | .094 | .390 |
| 2 | .169 | .497 |
| 3 | .136 | .475 |
| 4 | .108 | .384 |
| 5 | .096 | .353 |
| 6 | .153 | .384 |
| 7 | .098 | .343 |
| 8 | .250 | .670 |
| 9 | .267 | .538 |
| 10 | .277 | .508 |
| 11 | .293 | .491 |
| 12 | .295 | .460 |
| 13 | .276 | .458 |
| 14 | .265 | .456 |
| 15 | .210 | .451 |
| 16 | .264 | .438 |
| 17 | .222 | .407 |
| 18 | .205 | .398 |
| 19 | .204 | .388 |
| 20 | .226 | .332 |
| 21 | .220 | .326 |
| 22 | .242 | .321 |
| 23 | .317 | .323 |
| 24 | .318 | .348 |
| 25 | .335 | .440 |
| 26 | .337 | .339 |
| 27 | .345 | .612 |
| 28 | .346 | .327 |
| 29 | .349 | .386 |
| 30 | .353 | .375 |

Table A-3

Item difficulty and discrimination indices for Measurement Test 2

| Item No. | Difficulty<br>(p) | Discrimination<br>($r_b$) |
|---|---|---|
| 1 | .305 | .700 |
| 2 | .389 | .433 |
| 3 | .299 | .403 |
| 4 | .374 | .409 |
| 5 | .365 | .353 |
| 6 | .386 | .349 |
| 7 | .397 | .349 |
| 8 | .361 | .306 |
| 9 | .398 | .396 |
| 10 | .471 | .385 |
| 11 | .488 | .348 |
| 12 | .445 | .333 |
| 13 | .458 | .730 |
| 14 | .458 | .695 |
| 15 | .458 | .637 |
| 16 | .482 | .603 |
| 17 | .458 | .612 |
| 18 | .458 | .611 |
| 19 | .447 | .553 |
| 20 | .557 | .398 |
| 21 | .537 | .398 |
| 22 | .507 | .396 |
| 23 | .512 | .379 |
| 24 | .585 | .369 |
| 25 | .538 | .371 |
| 26 | .554 | .373 |
| 27 | .553 | .354 |
| 28 | .550 | .341 |
| 29 | .506 | .331 |
| 30 | .542 | .307 |

Table A-4

Item difficulty and discrimination indices for Measurement Test 3

| Item No. | Difficulty (p) | Discrimination ($r_b$) |
|---|---|---|
| 1 | .687 | .604 |
| 2 | .695 | .403 |
| 3 | .677 | .540 |
| 4 | .698 | .500 |
| 5 | .681 | .464 |
| 6 | .686 | .474 |
| 7 | .667 | .320 |
| 8 | .628 | .302 |
| 9 | .749 | .610 |
| 10 | .693 | .557 |
| 11 | .793 | .555 |
| 12 | .795 | .581 |
| 13 | .783 | .504 |
| 14 | .720 | .496 |
| 15 | .721 | .495 |
| 16 | .733 | .490 |
| 17 | .728 | .464 |
| 18 | .719 | .457 |
| 19 | .726 | .462 |
| 20 | .708 | .461 |
| 21 | .708 | .457 |
| 22 | .699 | .485 |
| 23 | .759 | .441 |
| 24 | .754 | .438 |
| 25 | .766 | .424 |
| 26 | .746 | .410 |
| 27 | .791 | .373 |
| 28 | .757 | .386 |
| 29 | .759 | .385 |
| 30 | .788 | .377 |

Table A-5

Item difficulty and discrmination indices for Measurement Test 4

| Item No. | Difficulty<br>(p) | Discrimination<br>($r_b$) |
|---|---|---|
| 1 | .827 | .579 |
| 2 | .843 | .551 |
| 3 | .811 | .550 |
| 4 | .895 | .524 |
| 5 | .806 | .508 |
| 6 | .857 | .487 |
| 7 | .807 | .458 |
| 8 | .875 | .430 |
| 9 | .850 | .405 |
| 10 | .813 | .402 |
| 11 | .831 | .382 |
| 12 | .884 | .367 |
| 13 | .885 | .376 |
| 14 | .866 | .376 |
| 15 | .890 | .367 |
| 16 | .904 | .506 |
| 17 | .911 | .537 |
| 18 | .921 | .565 |
| 19 | .926 | .410 |
| 20 | .928 | .366 |
| 21 | .942 | .385 |
| 22 | .948 | .447 |
| 23 | .958 | .487 |
| 24 | .963 | .560 |
| 25 | .921 | .751 |
| 26 | .932 | .776 |
| 27 | .937 | .693 |
| 28 | .943 | .699 |
| 29 | .953 | .660 |
| 30 | .958 | .710 |

## Appendix B

## Item Specifications for Linear Test

Table B-1

Item difficulty and discrimination indices for the linear test

| Item No. | Difficulty ($p$) | Discrimination ($r_b$) |
|---|---|---|
| 1 | .661 | .434 |
| 2 | .656 | .543 |
| 3 | .659 | .490 |
| 4 | .660 | .472 |
| 5 | .646 | .520 |
| 6 | .646 | .477 |
| 7 | .651 | .531 |
| 8 | .640 | .494 |
| 9 | .634 | .534 |
| 10 | .634 | .503 |
| 11 | .623 | .456 |
| 12 | .610 | .518 |
| 13 | .608 | .371 |
| 14 | .613 | .320 |
| 15 | .607 | .516 |
| 16 | .615 | .315 |
| 17 | .604 | .427 |
| 18 | .602 | .538 |
| 19 | .590 | .433 |
| 20 | .560 | .474 |
| 21 | .557 | .448 |
| 22 | .559 | .501 |
| 23 | .559 | .527 |
| 24 | .549 | .496 |
| 25 | .542 | .451 |
| 26 | .539 | .531 |
| 27 | .542 | .490 |
| 28 | .529 | .424 |
| 29 | .530 | .500 |
| 30 | .514 | .448 |
| 31 | .500 | .519 |
| 32 | .506 | .428 |
| 33 | .449 | .520 |
| 34 | .470 | .400 |
| 35 | .463 | .537 |
| 36 | .439 | .466 |
| 37 | .434 | .451 |
| 38 | .420 | .437 |
| 39 | .419 | .482 |
| 40 | .406 | .489 |

# DISTRIBUTION LIST

Navy

4 Dr. Marshall J. Farr, Director
Personnel & Training Research Programs
Office of Naval Research
Arlington, VA 22217

1 Director
ONR Branch Office
495 Summer Street
Boston, MA 02210
ATTN: C. M. Harsh

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

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

1 Office of Naval Research
Area Office
207 West 24th Street
New York, NY 10011

6 Director
Naval Research Laboratory
Code 2627
Washington, DC 20390

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

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

1 Chief of Naval Technical Training
Naval Air Station Memphis (75)
Millington, TN 38054
ATTN: Dr. G. D. Mayo

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

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

1 Commander
Naval Air Reserve
Naval Air Station
Glenview, IL 60026

1 Commander
Naval Air Systems Command
Department of the Navy
AIR-413C
Washington, DC 20360

1 Mr. Lee Miller (AIR 413E)
Naval Air Systems Command
5600 Columbia Pike
Falls Church, VA 22042

1 Dr. Harold Booher
NAVAIR 415C
Naval Air Systems Command
5600 Columbia Pike
Falls Church, VA 22042

1 CAPT John F. Riley, USN
Commanding Officer
U.S. Naval Amphibious School
Coronado, CA 92155

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

1 Dr. Richard J. Niehaus
Office of Civilian Manpower Management
Code 06A
Department of the Navy
Washington, DC 20390

1 CDR Richard L. Martin, USN
COMFAIRMIRAMAR F-14
NAS Miramar, CA 92145

1 Research Director, Code 06
Research and Evaluation Department
U.S. Naval Examining Center
Great Lakes, IL 60088
ATTN: C. S. Winiewicz

1 Chief
Bureau of Medicine and Surgery
Code 413
Washington, DC 20372

1 Program Coordinator
Bureau of Medicine and Surgery (Code 71G)
Department of the Navy
Washington, DC 20372

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

1 Technical Reference Library
Naval Medical Research Institute
National Naval Medical Center
Bethesda, MD 20014

1 Chief
Bureau of Medicine and Surgery
Research Division (Code 713)
Department of the Navy
Washington, DC 20372

1 Dr. John J. Collins
Chief of Naval Operations (OP-987F)
Department of the Navy
Washington, DC 20350

1 Technical Library (Pers-11B)
Bureau of Naval Personnel
Department of the Navy
Washington, DC 20360

1 Head, Personnel Measurement Staff
Capital Area Personnel Office
Ballston Tower #2, Room 1204
801 N. Randolph Street
Arlington, VA 22203

1 Dr. James J. Regan, Technical Director
Navy Personnel Research
and Development Center
San Diego, CA 92152

1 Mr. E. P. Somer
Navy Personnel Research
and Development Center
San Diego, CA 92152

1 Dr. Norman Abrahams
Navy Personnel Research
and Development Center
San Diego, CA 92152

1 Dr. Bernard Rimland
Navy Personnel Research
and Development Center
San Diego, CA 92152

1 Commanding Officer
Navy Personnel Research
and Development Center
San Diego, CA 92152

1 Superintendent
Naval Postgraduate School
Monterey, CA 92940
ATTN: Library (Code 2124)

1 Mr. George N. Graine
Naval Ship Systems Command
(SHIPS 03H)
Department of the Navy
Washington, DC 20360

1 Technical Library
Naval Ship Systems Command
National Center, Building 3
Room 3S08
Washington, DC 20360

1 Commanding Officer
Service School Command
U.S. Naval Training Center
San Diego, CA 92133
ATTN: Code 303

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

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

1 CDR Fred Richardson
Navy Recruiting Command
BCT #2, Room 215
Washington, DC 20370

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

1 Dr. H. Wallace Sinaiko
c/o Office of Naval Research (Code 450)
Psychological Sciences Division
Arlington, VA 22217

1 Dr. Martin F. Wiskoff
Navy Personnel Research
and Development Center
San Diego, CA 92152

### Army

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

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

1 Director of Research
U.S. Army Armor Human Research Unit
ATTN: Library
Building 2422 Morade Street
Fort Knox, KY 40121

1 U.S. Army Research Institute for the
Behavioral and Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

1 Commanding Officer
ATTN: LTC Montgomery
USACDC - PASA
Ft. Benjamin Harrison, IN 46249

1 Commandant
United States Army Infantry School
ATTN: ATSIN-H
Fort Benning, GA 31905

1 U.S. Army Research Institute
Commonwealth Building, Room 239
1300 Wilson Boulevard
Arlington, VA 22209
ATTN: Dr. R. Dusek

1 Mr. Edmund F. Fuchs
U.S. Army Research Institute
1300 Wilson Boulevard
Arlington, VA 22209

1 Commander
U.S. Theater Army Support Command,
Europe
ATTN: Asst. DCSPER (Education)
APO New York 09058

1 Dr. Stanley L. Cohen
Work Unit Area Leader
Organizational Development Work Unit
Army Research Institute for Behavioral
and Social Science
1300 Wilson Boulevard
Arlington, VA 22209

### Air Force

1 Headquarters, U.S. Air Force
Chief, Personnel Research and Analysis
Division (AF/DPSY)
Washington, DC 20330

1 Research and Analysis Division
AF/DPXYR Room 4C200
Washington, DC 20330

1 AFHRL/AS (Dr. G. A. Eckstrand
Wright-Patterson AFB
Ohio 45433

1 AFHRL/MD
701 Prince Street
Room 200
Alexandria, VA 22314

1 Dr. Robert A. Bottenberg
AFHRL/PES
Lackland AFB, TX 78236

1 Personnel Research Division
AFHRL
Lackland Air Force Base
Texas 78236

1 AFOSR(NL)
1400 Wilson Boulevard
Arlington, VA 22209

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

1 CAPT Jack Thorpe, USAF
Department of Psychology
Bowling Green State University
Bowling Green, OH 43403

Marine Corps

1 Commandant, Marine Corps
Code A01M-2
Washington, DC 20380

1 COL George Caridakis
Director, Office of Manpower Utilization
Headquarters, Marine Corps (A01H)
MCB
Quantico, VA 22134

1 Dr. A. L. Slafkosky
Scientific Advisor (Code Ax)
Commandant of the Marine Corps
Washington, DC 20380

1 Mr. E. A. Dover
Manpower Measurement Unit (Code A01M-2)
Arlington Annex, Room 2413
Arlington, VA 20370

Coast Guard

1 Mr. Joseph J. Cowan, Chief
Psychological Research Branch (P-1)
U.S. Coast Guard Headquarters
400 Seventh Street, SW
Washington, DC 20590

Other DOD

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

1 Mr. Helga Yeich, Director
Program Management, Defense Advanced
Research Projects Agency
1400 Wilson Boulevard
Arlington, VA 22209

1 Dr. Ralph R. Canter
Director for Manpower Research
Office of Secretary of Defense
The Pentagon, Room 3C980
Washington, DC 20301

Other Government

1 Dr. Lorraine D. Eyde
Personnel Research and Development Center
U.S. Civil Service Commission, Room 3458
1900 E. Street, N.W.
Washington, DC 20415

1 Dr. Vern Urry
Personnel Research and Development
Center
U.S. Civil Service Commission
Washington, DC 20415

Miscellaneous

1 Dr. Scarvia Anderson
Executive Director for Special Development
Educational Testing Service
Princeton, NJ 08540

1 Dr. Richard C. Atkinson
Stanford University
Department of Psychology
Stanford, CA 94305

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

1 Mr. H. Dean Brown
Stanford Research Institute
333 Ravenswood Avenue
Menlo Park, CA 94025

1 Mr. Michael W. Brown
Operations Research, Inc.
1400 Spring Street
Silver Spring, MD 20910

1 Dr. Ronald P. Carver
American Institutes for Research
8555 Sixteenth Street
Silver Spring, MD 20910

1 Century Research Corporation
4113 Lee Highway
Arlington, VA 22207

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

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

1 Dr. Norman R. Dixon
Associate Professor of Higher Education
University of Pittsburgh
617 Cathedral of Learning
Pittsburgh, PA 15213

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

1 Dr. Marvin D. Dunnette
University of Minnesota
Department of Psychology
N492 Elliott Hall
Minneapolis, MN 55455

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

1 Dr. Edwin A. Fleishman
American Institutes for Research
8555 Sixteenth Street
Silver Spring, MD 20910

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

1 Dr. Albert S. Glickman
American Institutes for Research
8555 Sixteenth Street
Silver Spring, MD 20910

1 Dr. Duncan N. Hansen
Florida State University
Center for Computer-Assisted Instruction
Tallahassee, FL 32306

1 Dr. Harry H. Harman
Educational Testing Service
Division of Analytical Studies and Services
Princeton, NJ 08540

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

1 Dr. M. D. Havron
Human Sciences Research, Inc.
Westgate Industrial Park
7710 Old Springhouse Road
McLean, VA 22101

1 Human Resources Research Organization
Division #3
P.O. Box 5787
Presidio of Monterey, CA 93940

1 Human Resources Research Organization
Division #4, Infantry
P.O. Box 2086
Fort Benning, GA 31905

1 Human Resources Research Organization
Division #5, Air Defense
P.O. Box 6057
Fort Bliss, TX 79916

1 Human Resources Research Organization
Division #6, Library
P.O. Box 428
Fort Rucker, AL 36360

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

1 Dr. Norman J. Johnson
Carnegie-Mellon University
School of Urban and Public Affairs
Pittsburgh, PA 15213

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

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

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

1 Mr. Edmond Marks
109 Grange Building
Pennsylvania State University
University Park, PA 16802

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

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

1 Dr. Robert D. Pritchard
Assistant Professor of Psychology
Purdue University
Lafayette, IN 47907

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

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

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

1 Dr. Benjamin Schneider
University of Maryland
Department of Psychology
College Park, MD 20742

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

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

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

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

1 Dr. Charles A. Ullmann
Director, Behavioral Sciences Studies
Information Concepts Incorporated
1701 No. Ft. Myer Drive
Arlington, VA 22209

1 Dr. H. Peter Dachler
University of Maryland
Department of Psychology
College Park, MD 20742