DOCUMENT RESUME

ED 077 933 TM 002 728

AUTHOR Weiss, David J.; Betz, Nancy E.

TITLE Ability Measurement: Conventional or Adaptive?
INSTITUTION Minnesota Univ., Minneapolis. Dept. of Psychology.
SPONS AGENCY Office of Naval Research, Washington, D.C. Personnel and Training Research Programs Office.
REPORT NO RR-73-1
PUB DATE Feb 73
NOTE 77p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Ability; Comparative Analysis; Group Tests; Individual Tests; *Literature Reviews; *Measurement Instruments; Psychometrics; *Testing; Test Reliability; Test Validity

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. (Author)

TM

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

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)	2a. REPORT SECURITY CLASSIFICATION
University of Minnesota Department of Psychology	unclassified
	2b. GROUP

3. REPORT TITLE

Ability Measurement: Conventional or Adaptive?

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

Technical Report

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

David J. Weiss and Nancy E. Betz

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO OF REFS
February 1973	70	136

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
N00014-67-A-0113-0029	Research Report 73-1 Psychometric Methods Program
b. PROJECT NO. 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
	Personnel and Training Research Programs, Office of Naval Research

13. 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.

DD FORM 1 NOV 65 1473

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						
paper and pencil testing						
two-stage tests						
Bayesian testing						
reliability						
validity						
automated testing						

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 Test 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 as if 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 difficulty 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 $n+1^{th}$ 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 $n+1^{th}$ 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, 196).

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.

Non-Bayesian strategies. 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 simulation 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 sting 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 administrators; 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 $n+1^{th}$ 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.

Termination rules. 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 time-shared 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 gameformat 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. The use 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. 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. 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, P. 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.

DISTRIBUTION LIST

NAVY

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

1 Director
ONR Branch Office
495 Summer Street
Boston, MA 02210

1 Director
ONR Branch Office
1030 East Green Street
Pasadena, CA 91101

1 Director
ONR Branch Office
536 South Clark Street
Chicago, IL 60605

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

1 Office of Naval Research
Area Office
1076 Mission Street
San Francisco, CA 94103

1 Commander
Operational Test and Evaluation Force
U.S. Naval Base
Norfolk, VA 23511

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 Allen E. McMichael

1 Chief
Bureau of Medicine and Surgery
Code 513
Washington, DC 20390

1 Chief
Bureau of Medicine and Surgery
Research Division (Code 713)
Department of the Navy
Washington, DC 20390

1 Commandant of the Marine Corps
(Code A01M)
Washington, DC 20380

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

1 Commander
Submarine Development Group Two
Fleet Post Office
New York, NY 09501

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

1 Commanding Officer
Naval Personnel and Training
Research Laboratory
San Diego, CA 92152

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

1 Program Coordinator
Bureau of Medicine and Surgery
(Code 71G)
Department of the Navy
Washington, DC 20390

1 Research Director, Code 06
Research and Evaluation Department
U. S. Naval Examining Center
Building 2711 - Green Bay Area
Great Lakes, IL 60088
ATTN: C. S. Winiewicz

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

1 Technical Director
Naval Personnel Research and
Development Laboratory
Washington Navy Yard
Building 200
Washington, DC 20390

1 Technical Director
Personnel Research Division
Bureau of Naval Personnel
Washington, DC 20370

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

1 Technical Library
Naval Ship Systems Command
National Center
Building 3 Room 3
S-08
Washington, DC 20360

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

1 COL George Caridakis
Director, Office of Manpower
Utilization
Headquarters, Marine Corps
(AO1H)
MCB
Quantico, VA 22134

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

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

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

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

1 Dr. James J. Regan
Code 55
Naval Training Device Center
Orlando, FL 32813

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

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

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

ARMY

1 Behavioral Sciences Division
Office of Chief of Research and Development
Department of the Army
Washington, DC 20310

1. U.S. Army Behavior and Systems Research Laboratory
Rosslyn Commonwealth Building, Room 239
1300 Wilson Boulevard
Arlington, VA 22209

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

1 COMMANDANT
U.S. Army Adjutant General School
Fort Benjamin Harrison, IN 42616
ATTN: ATSAG-EA

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

1 Director
Behavioral Sciences Laboratory
U.S. Army Research Institute of Environmental Medicine
Natick, MA 01760

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

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

1 Mr. Edmund Fuchs
BESRL
Commonwealth Building, Room 239
1320 Wilson Boulevard
Arlington, VA 22209

AIR FORCE

1 AFHRL (TR/Dr. G. A. Eckstrand)
Wright-Patterson Air Force Base
Ohio 45433

1 AFHRL (TRT/Dr. Ross L. Morgan)
Wright-Patterson Air Force Base
Ohio 45433

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

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

1 COMMANDANT
USAF School of Aerospace Medicine
ATTN: Aeromedical Library (SCL-4)
Brooks AFB, TX 78235

1 Personnel Research Division
AFHRL
Lackland Air Force Base
San Antonio, TX 78236

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

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

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

DOD

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

1 Dr. Ralph R. Canter
Director for Manpower Research
Office of Secretary of Defense
The Pentagon, Room 30980
Washington, DC 20301

1 Dr. Charles Ullman
Chief of Counseling, Training
Programs
OSD(M&RA)
The Pentagon, Room 2C252
Washington, DC 20301

OTHER GOVERNMENT

1 Dr. Alvin E. Goins, Chief
Personality and Cognition Research
Section
Behavioral Sciences Research Branch
National Institute of Mental Health
5600 Fishers Lane
Rockville, MD 20852

1 Dr. Andrew R. Molnar
Computer Innovation in Education
Section
Office of Computing Activities
National Science Foundation
Washington, DC 20550

1 Dr. Lorraine D. Eyde
Bureau of Intergovernmental Personnel
Programs
Room 2519
U.S. Civil Service Commission
1900 E. Street, NW
Washington, DC 20415

1 Office of Computer Information
Center for Computer Sciences and
Technology
National Bureau of Standards
Washington, DC 20234

MISCELLANEOUS

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

1 Professor John Annett
The Open University
Walton teale, BLETCHLEY
Bucks, ENGLAND

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

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

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

1 Dr. Rene V. Dawis
Department of Psychology
324 Elliott Hall
University of Minnesota
Minneapolis, MN 55455

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

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

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

1 Mr. Paul P. Foley
Naval Personnel Research and
Development Laboratory
Washington Navy Yard
Washington, DC 20390

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

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

1 Dr. Bert Green
Department of Psychology
Johns Hopkins University
Baltimore, MD 21218

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

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
Post Office Box 5787
Presidio of Monterey, CA 93940

1 Human Resources Research Organization
Division #4, Infantry
Post Office Box 2086
Fort Benning, GA 31905

1 Human Resources Research Organization
Division #5, Air Defense
Post Office Box 6057
Fort Bliss, TX 79916

1 Library
HumRRO Division Number 6
P. O. Box 428
Fort Rucker, AL 36360

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

1 Dr. Roger A. Kaufman
Graduate School of Human Behavior
U.S. International University
8655 E. Pomerada Road

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

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

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

1 Dr. Stanley M. Nealy
Department of Psychology
Colorado State University
Fort Collins, CO 80521

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 Psychological Abstracts
American Psychological Association
1200 Seventeenth Street, NW
Washington, DC 20036

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 Calif.
3717 South Grand
Los Angeles, Calif. 90007

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

1 Dr. George E. Rowland
Rowland and Company, Inc.
Post Office Box 61
Haddonfield, NJ 08033

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

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

1 Dr. Henry Solomon
George Washington University
Department of Economics
Washington, DC 20006

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

1 LCOL Austin W. Kibler, Director
Human Resources Research Office
ARPA
1400 Wilson Boulevard
Arlington, VA 22209

1 Century Research Corporation
4113 Lee Highway
Arlington, VA 22207

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

1 Dr. Kenneth E. Young
Vice President
American College Testing Program
Suite 340
One DuPont Circle, N. W.
Washington, DC 20036

1 Mr. Emanuel P. Somer
Head, Motivational and Survey
Research Division
Psychological Research Department
Naval Personnel R&D Laboratory
Washington Navy Yard
Washington, DC 20374