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

Nancy E. Betz

and

David J. Weiss

Research Report 73-4

Psychometric Methods Program
Department of Psychology
University of Minnesota

October 1973

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

Security Classification

DOCUMENT CONTROL DATA - R & D

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

1. ORIGINATING ACTIVITY *(Corporate author)*

Department of Psychology
University of Minnesota

2a. REPORT SECURITY CLASSIFICATION

Unclassified

2b. GROUP

3. REPORT TITLE

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

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

Technical Report

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

Nancy E. Betz and David J. Weiss

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
October 1973	49	35

8a. CONTRACT OR GRANT NO.

N00014-67-A-0113-0029

b. PROJECT NO.

NR150-343

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)

Research Report 73-4
Psychometric Methods Program

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

10. DISTRIBUTION STATEMENT

Approved for public release; distribution unlimited

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY

Personnel & Training Research Programs
Office of Naval Research

13. ABSTRACT

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

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

Contents

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

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

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

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

Results showed the measurement tests to be more reliable in the groups for which they were intended than conventional broad-range tests. Predictive validities of the measurement tests, using grade point averages as the criterion, were slightly higher than those of the conventional tests. Their

data also showed, however, that about 20% of the testees would have been misclassified, or routed to an inappropriate measurement test.

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

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

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

Scores on the two-stage tests were initially determined by scaling the measurement tests using linear regression weights to predict the total score on the parent test. A

later study (Linn et al., 1969) added the routing score information to the scaled measurement test score.

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

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

Lord (1971d) presents results from theoretical studies of two-stage testing procedures. All of his analyses were based on the mathematics of item characteristic curve theory and the following assumptions: 1) a fixed number of items administered to each examinee, 2) dichotomous (right-wrong) scoring, 3) normal ogive item characteristic curves, 4) a unidimensional set of items, 5) all items of equal discriminations, 6) peaked routing and measurement tests (i.e., all items in each subtest were of the same difficulty), and 7) linear (i.e., non-branched) routing and measurement tests. Lord studied about 200 different strategies, varying the total number of items (15 or 60), the number of alternative measurement tests, the cutting points for assignment to the second-stage tests, methods of scoring both the routing test and the entire two-stage procedure, and whether or not random

guessing was assumed (for a 5-choice item, within the 60-item tests only). Lord compared each two-stage strategy with a peaked conventional test of equivalent length in terms of information functions, which indicate the relative numbers of items required to achieve equivalent precision of measurement. Precision can be defined as the capability of responses to a set of test items to accurately represent the "true ability" of hypothetical individuals.

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

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

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

Lord's theoretical study of two-stage testing procedures, based on the notion that a short routing test can be used to find the optimal peaked measurement test for any given individual, as well as the studies of Angoff and Huddleston (1958)

and Cleary et al. (1968a,b; Linn et al., 1969) show considerable potential for two-stage tests, in terms of increases in internal consistency reliability, validity, and precision of measurement. However, only Angoff and Huddleston's was an empirical study, and even this study was not able to account for the effects of actual routing. The purpose of the present study, then, was to begin an empirical evaluation of two-stage testing procedures; the study involved the development, computer-controlled administration, and comparison of a two-stage test and a peaked conventional test.

METHOD

Design

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

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

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

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

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

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

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

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

variability than scores from peaked conventional tests. In addition, the score distributions were examined to determine whether the tests yielded skewed, rectangular, peaked, or non-unimodal distributions.

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

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

Lord's (1971d) theoretical demonstration that the precision of measurement of two-stage tests was nearly constant over the whole ability range implies fewer random factors in the ordering of individuals in two-stage tests than in conventional tests. In conventional tests, which are most precise at average ability levels, scores of individuals near the extremes of ability will be highly affected by random errors, and the ordering of such individuals will

be determined in large part by random factors. Because of the more nearly constant precision of two-stage tests, the scores for individuals at all levels of the ability distribution are more likely to be based largely on underlying ability rather than on random factors; two-stage tests should thus yield higher test-retest stability coefficients than conventional tests. One complicating factor, however, involves differential memory effects. A subject re-tested on the conventional test will repeat the same set of items. A subject retested on the two-stage test will take the same set of 10 routing items but may take an entirely different set of 30 measurement test items if he is routed differently the second time. In comparing the stability, then, of two-stage and conventional tests, it was necessary to account for the differential effects of memory.

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

Test Development

Item Pool

The item pool used to construct the adaptive and conventional tests of verbal ability consisted of 5-alternative multiple-choice vocabulary items. The items were normed on a large group of college students, and item statistics of

difficulty (proportion correct) and discrimination (biseral correlation with total score) were obtained. Using a biserial correlation of at least .30 as a selection criterion, 369 items were available for use in constructing the three tests to be administered in the first study. Table 1 describes the available item pool as a cross-classification of levels of item difficulty and biserial correlation coefficient and shows the number of items available in each cell of the cross-tabulation. It may be noted that the pool consisted of considerably more very easy than very difficult items, and that the more highly discriminating items occurred at the easier levels of difficulty.

Two-stage Test

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

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

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

Table 1

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

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

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

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

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

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

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

Table 2

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

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

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

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

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

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

The resulting characteristics of the four measurement tests are summarized in Table 2. It may be noted that the mean item difficulties of tests 1 and 4 were slightly different from the desired values; this was due to the necessity of taking item discrimination as well as item difficulty into account. However, the resulting values of $a(b_2 - b)$, which were +1.09, +.39, -.40, and -1.13, were good approximations to the values specified beforehand. As with the routing test, item difficulties of each of the measurement tests were normally distributed around the mean value. Also, the mean biserial correlations for the two most difficult measurement tests were lower than those for the two easier tests. This was due to the relative scarcity of difficult items having high biserial coefficients as was indicated in Table 1. And while the mean biserial levels were relatively close to the .45 value desired, the standard

deviation and range of these values show that it was not possible to construct equi-discriminating tests using the available item pool within the limitations of the research design (i.e., the construction of several non-overlapping tests). Appendix Tables A-2 through A-5 give the characteristics of each of the 30 items in each measurement test in terms of p-values and biserial coefficients.

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

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

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

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

Lord (1971d) admits that there is no uniquely good way to weight the subtest $\hat{\theta}$'s. He computed variance weights, but a preliminary examination of the results of applying his weighting formula to the two-stage data from this study showed some non-monotonicity in the relationship between the number right obtained on the measurement test and the

total test $\hat{\theta}$ for people who obtained the same routing score. Therefore, rather than using the variance weights, each subtest $\hat{\theta}$ was weighted according to the number of items on which it was based; the resulting total score estimates were then strictly monotonically related to the actual number correct on the measurement test, given the same routing score. The ability estimate used in this study, then, was defined by the following equation:

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

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

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

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

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

Administration and Subjects

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

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

Analysis of Data

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

Table 3

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

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

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

Characteristics of Score Distributions

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

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

Reliability

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

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

---

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

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

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

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

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

Additional Analyses

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

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

RESULTS

Comparison of Two-stage and Linear Tests on Psychometric Characteristics

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

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

Table 4

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

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

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

Table 5

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

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

*significant at $p < .02$

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

Reliability

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

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

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

Table 6

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

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

Table 7

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

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

Table 8

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

	Time 1				Time 2			
Test	N	Mean	S.D.	S.D./No. of items	N	Mean	S.D.	S.D./No. of items
Routing	214	6.78	2.31	.23	178	7.18	2.28	.23
Measurement								
1	91	12.98	5.28	.18	93	13.34	5.25	.18
2	61	15.38	4.51	.15	41	14.10	4.28	.14
3	39	19.13	3.33	.11	28	20.79	3.48	.12
4	23	23.39	4.81	.16	16	24.19	3.82	.13
Linear	110	24.13	8.29	.21	85	25.67	8.32	.21

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

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

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

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

Table 9

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

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

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

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

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

Relationships between Linear and Two-stage Scores

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

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

Table 10

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

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

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

Comparison of Norming and Testing Item Statistics

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

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

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

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

Table 11

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

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

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

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

Item discriminations. Table 11 presents the means and standard deviations of item discrimination values (biserial correlation with number correct) as obtained from the administration of the tests. A comparison of these values with the norming values as presented in Table 2 shows that the testing mean item discrimination values for the linear and routing test were higher than the corresponding norming values; the mean biserials of the linear test items were .47 from the norming studies but .56 and .58 from the test and retest, and the routing test increased from a mean discrimination of .57 in norming to .67 and .69. In contrast, the only measurement test to show higher item discrimination values on both test and retest was test 1, the most difficult test, whose means were .42 in norming but .49 and .46 on test and retest. The items in tests 2 and 3 were less discriminating in testing than they had been in norming, and those in test 4 were more discriminating on the first test but less discriminating on the retest. Further, the standard deviations of the item discrimination values were again larger in testing than they had been in norming. The items in test 4 especially showed much greater variability in their discriminating power.

The substantial changes that were found in both the level and variability of item discriminating power were probably a factor in the rather poor internal consistency reliability of the measurement tests and suggest that item statistics derived from norming samples composed of one range of ability levels may be inappropriate when applied to a group composed of a different range of ability levels.

Additional Characteristics of the Two-stage Test

The results thus far have suggested certain problems with the two-stage test. Three of the four measurement tests were not of optimal difficulty for the groups of subjects taking them, and the item discrimination values of the measurement tests tended to be both lower and more variable in actual two-stage testing than they had been in norming. Thus, the two-stage test was further examined to evaluate the degree to which it met its major objective. That is, the two-stage test was analyzed to determine whether the "routing" test assigned members of a group of individuals varying rather widely in ability to longer "measurement" tests such that each measurement test was essentially "peaked" at the mean ability of a far more homogeneous group of subjects and was thus more appropriate to their level of ability than would be a test designed to measure the full range of ability within the larger group.

In first examining the characteristics of the 10-item routing test, it was found that the mean number correct was 6.78 on the first test and 7.18 on the retest (see Table 8). These high mean scores were close to expectation because the test was constructed to be somewhat easier than the median ability with chance success accounted for (p=.60). However, on both test and retest, the distribution of routing test scores showed a significant degree of negative skew, indicating a predominance of high scores (7 to 10 correct).

The high and significantly skewed routing scores, coupled with the score intervals selected for assignment to measurement tests (0-3, 4-5, 6-7, and 8-10), meant that a majority of the testees were assigned to the two most difficult measurement tests (tests 1 and 2). Table 12 summarizes data on the number and percentage of the total group assigned to each measurement test and the mean and standard deviation of the number correct scores obtained by each of these subgroups.

The data in Table 12 show several deficiencies of the two-stage test used in this study. First, the imbalance in the numbers of testees taking the individual measurement tests is obvious and consistent; roughly half of the total group took the most difficult test on both test and retest, whereas only about one-tenth of the group took the easiest test. Although the percentages taking each test time 1 and time 2 are fairly comparable, there was a tendency for the imbalance to be even more pronounced on the retest.

Table 12

Number and percentage of total group assigned to each measurement test and mean and standard deviation of number correct (of 30 possible) for each test

Measurement test	Score range on routing test	Time 1				Time 2			
				Number correct				Number correct	
		N	%	Mean	S.D.	N	%	Mean	S.D.
1	8-10	91	42.5	12.98	5.28	93	52.2	13.34	5.25
2	6-7	61	28.5	15.38	4.51	41	23.0	14.10	4.28
3	4-5	39	18.2	19.13	3.33	28	15.7	20.79	3.48
4	0-3	23	10.7	23.39	4.81	16	9.0	24.19	3.82

Second, as was pointed out in the section on reliability, the tests were not of optimal difficulty for those groups of individuals taking them. The most appropriate mean item difficulty would be around p=.60, meaning that the desired mean number correct on each measurement test would be about 18. As Table 12 shows, however, the two most difficult tests were too difficult (mean total scores of 12.98 and 15.38 respectively) for the average subject taking them, and the two easier tests were too easy (means of 19.13 and 23.39 respectively). These results and the findings of the rather low number-correct score variability of the measurement tests, as shown in Table 8 and discussed in the reliability section, suggest that the total group was more homogeneous in ability than expected. If the cutting scores for assignment to measurement tests had been set higher, e.g., 0-4, 5-6, 7-8, and 9-10, the two most difficult measurement tests would probably have been more appropriate, but the placement of higher ability subjects into the easier tests would have made these two tests even easier, and thus more inappropriate for many of the individuals assigned to them, than they were using the score intervals selected for this two-stage test.

Misclassification. A different approach to the evaluation of the appropriateness of assignment to measurement tests was to identify the extent to which particular individuals were classified into inappropriate tests. Defining misclassified individuals as those who obtained perfect scores (e.g., all 30 items correct), indicating that the test was too easy, or scores at or below chance (i.e., scores of 6 or less correct), indicating that the test was too difficult, there were 9 or 4.2% misclassifications on the first test and 9 or 5.0% on the retest. All 18 misclassifications were the result of scores at or below chance on the most difficult measurement test, thus providing additional evidence that this test was too difficult for many individuals routed to it. However, the 4 to 5% misclassification rate obtained here was a considerable improvement over the 20% rates obtained in the studies of Angoff and Huddleston (1958) and Cleary et al. (1968a,b), although this may be due in part to different criteria of misclassification. Thus, although the measurement tests were not optimal for the groups taking them, few individuals took a test which was highly inappropriate.

CONCLUSIONS AND IMPLICATIONS

Considering that the two-stage test used in this study had some deficiencies, the findings of the study were generally favorable to the continued exploration of two-stage testing

procedures. The two-stage test, scored using a variation of the method used in Lord's (1971d) theoretical study, yielded scores which were normally distributed and utilized a consistently higher proportion of the available score range than did the linear test. In other empirical studies of adaptive testing where the distribution of scores has been examined, a tendency toward badly skewed scores with definite bunching at the high end of the distribution has been found (Bayroff & Seeley, 1967; Bayroff, Thomas & Anderson, 1960; Seeley, Morton, & Anderson, 1962). Thus, the two-stage test constructed for this study yielded a better distribution of scores than has been found in most empirical studies of adaptive testing to date. The significantly flat distribution of linear test scores may have been a function of deviations from peakedness in its construction; a more peaked test might have yielded a more normal distribution of scores.

The findings regarding the reliability of the two-stage test were less clear. In terms of test-retest stability, the two-stage test scores were quite reliable ($r=.88$) over a mean interval of 5.5 weeks, essentially as stable as the linear test scores ($r=.89$). However, when the effect of memory of the items was equated for the two testing strategies, the two-stage scores were the more stable ($r=.93$). Thus, the two-stage test yielded 7.3% more stable variance than did the linear test of the same number of items and with the same potential for memory effects.

The relatively poor internal consistency reliability of the measurement tests, as compared to the high reliabilities of the routing test and the conventional linear test, was a finding in contrast to those of Angoff and Huddleston (1958) and was probably due to a combination of factors. First, the routing process created subgroups of individuals who were very homogeneous in ability. This was not an unexpected finding, especially given the relative homogeneity of ability in a college student population in comparison to that in a more general population. Further, even though increasing subgroup homogeneity decreases internal consistency, the purpose of the two-stage test is to do precisely that; by initially classifying a group of subjects as to ability, as the routing test does, it is possible to measure them using the most appropriate peaked measurement test. The best two-stage testing procedure would be one containing an infinite number of measurement tests, such that there would be a peaked test perfectly suited to each individual's ability. In this hypothetical mode of testing, there would be complete

homogeneity of ability within subgroups since each measurement test would be taken only by individuals with exactly equal ability. Thus, it is perhaps unrealistic to expect high internal consistency reliability from tests which function in this way.

In addition to the extreme subgroup homogeneity, the item difficulties of the measurement tests were not optimal for high reliability, and many of the items which had been highly discriminating in the norming studies were much less discriminating when administered to more homogeneous samples from the total group, thus reducing the internal consistency. Both of these inadequacies can be traced to the inappropriateness of traditional methods of determining item parameters for items to be used in adaptive testing. Only after administering a two-stage test to a defined group of individuals is it possible to determine how difficult and how discriminating the items will be for each subgroup of individuals formed; thus, selecting items for two-stage tests using traditional item parameters can at present be only an approximate procedure. Perhaps the construction of future two-stage tests should use item parameters derived from heterogeneous samples for selection of the routing test items but item parameters derived from more homogeneous subgroups of the total norming sample for the selection of items for each of the measurement tests. Alternatively, item parameters estimated using the techniques of modern test theory (e.g., Lord & Novick, 1968) might be appropriate if it can be shown that these parameters are independent of the range and level of ability in the groups on which they are determined.

The selection of score intervals for assignment to measurement tests is also a matter that needs further study. In this study, the score intervals selected were somewhat inappropriate, leading to an uneven distribution of testees among measurement tests. Although the measurement tests were more appropriate in difficulty for the groups taking them than a test peaked at the median total-group difficulty would be, they were still either somewhat too easy or somewhat too difficult for the groups taking them. However, few individuals were misclassified under the criteria used; the 5% rate of misclassification was a large improvement over the 20% rates of Angoff and Huddleston's (1958) and Cleary et al.'s (1968a,b; Linn et al., 1969) two-stage tests.

The relationship between the linear and two-stage test scores was relatively high (.84 and .80) and primarily linear. The nonlinearity that was found in the regression

of the linear scores on the two-stage scores on the first test seemed to be due to restriction in the lower score ranges of the linear test in comparison to the lack of range restriction in the two-stage scores. However, further analyses showed that the relationship between the two tests left about 20% of the reliable variance in the linear test scores and an unknown amount of reliable variance in the two-stage test scores unaccounted for.

A conventional linear test, however, should not be taken as a standard against which new methods of testing must be evaluated. Although a peaked conventional test provides probably the most accurate measurement for individuals whose ability level is near the group mean or the difficulty level at which the test is peaked, its accuracy becomes increasingly less as an individual's ability level deviates from the mean (Lord, 1970, 1971a,c,d). Adaptive tests, on the other hand, provide almost constant accuracy throughout the range of ability (Lord, 1970, 1971a,c,d). Thus, the relationship between the two-stage and linear tests can become meaningful only in the comparative context of indices of relationship between other adaptive strategies and the two-stage test, and indices of the extent to which the two-stage test and the linear test are found to predict a variety of relevant external criteria. Previous studies of two-stage and other adaptive testing strategies have found the adaptive tests to have higher relationships with external criteria than conventional tests of equivalent length (Angoff and Huddleston, 1958; Linn et al., 1969; Waters, 1964, 1970; Waters & Bayroff, 1971; see Weiss & Betz, 1973). No studies to date have examined the relationships between two or more adaptive tests. Thus, the validation of two-stage testing procedures depends on additional research in this area.

For further study of two-stage testing procedures, it should be possible to use the information gained in this study to select more optimal score intervals for assignment to measurement tests, to select more appropriate measurement test item difficulties, and to improve the internal consistency reliability by selecting items shown to be highly discriminating for particular subgroups as well as for the total group. A method of selecting the routing test score intervals that would probably be superior to rational or trial-and-error selection would be to compute each individual's latent ability estimate from the routing test ($\hat{\theta}_i$, as described in the scoring section) and to assign him to that measurement test whose mean difficulty in normal ogive parameter terms ("b" values) is closest to the estimate of his/her ability derived from the routing test.

However, the most obvious deficiency of two-stage testing procedures in general is that individuals may be routed to highly inappropriate measurement tests. A low ability individual may guess enough routing items correctly to place him in a measurement test that is too difficult. A higher ability individual confronted with a set of routing items that he is unable to answer correctly as a result of specific gaps in his knowledge or anxiety at the early stages of testing would be routed to a measurement test that is too easy.

One approach to this problem, of course, would be to lengthen the routing test. This approach, however, would undermine one advantage of two-stage testing, i.e., to arrive at an initial estimate of each individual's ability as quickly and efficiently as possible so that a larger set of items relevant to his/her ability may be administered. A more desirable approach would seem to be to include a recovery routine in the computer program controlling test administration. This routine would detect individuals who had apparently been misclassified after only a few measurement test items had been administered; for example, a chance score or a near-perfect score after 10 measurement test items had been administered would cause the individual to be re-routed into the next easier or next more difficult measurement test. The process could be repeated if following re-routing the individual was still wrongly classified. This procedure would mean that individuals would complete different total numbers of items depending on the ease or difficulty of correctly classifying them; thus, the number as well as the difficulty level of the items administered would be adapted to each individual.

Much empirical research remains to be done on two-stage testing procedures; if the information gained from previous empirical studies and the possibilities for improvements suggested by these studies can be fully utilized in subsequent research, it is likely that two-stage testing procedures will become valuable and practical alternatives to traditional testing procedures.

References

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

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

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

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

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

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

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

Cronbach, L. J. & Gleser, G. C. Psychological tests and personnel decisions. (2nd Ed.) Urbana: University of Illinois Press, 1965.

Cronbach, L. J. & Warrington, W. G. Efficiency of multiple-choice tests as a function of spread of item difficulties. Psychometrika, 1952, 17, 127-147.

Evans, R. N. A suggested use of sequential analysis in performance acceptance testing. Urbana: College of Education, University of Illinois, mimeo, 1953.

Guilford, J. P. Psychometric methods. New York: McGraw-Hill, 1954.

Hoyt, C. J. Test reliability estimated by analysis of variance. Psychometrika, 1941, 3, 153-160.

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

Linn, R. L., Rock, D. A., & Cleary, T. A. The development and evaluation of several programmed testing methods. Educational and Psychological Measurement, 1969, 29, 129-146.

Lord, F. M. The relation of the reliability of multiple-choice tests to the distribution of item difficulties. Psychometrika, 1952, 17, 181-194.

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

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

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

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

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

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

McNemar, Q. Psychological statistics. (4th ed.) New York: Wiley, 1969.

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

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

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

Statistical Research Group, Columbia University. Sequential analysis of statistical data, applications. New York: Columbia University Press, 1945.

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

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

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

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

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

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

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

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

Wood, R. Response-contingent testing. Review of Educational Research, 1973 (in press).

Appendix A

Item Specifications for Two-stage Test

Table A-1

Item difficulty and discrimination indices for the Routing Test

Item No.	Difficulty (p)	Discrimination (r_b)
1	.568	.708
2	.566	.653
3	.589	.563
4	.635	.608
5	.626	.552
6	.622	.552
7	.675	.566
8	.674	.554
9	.677	.547
10	.598	.430

Table A-2

Item difficulty and discrimination indices for Measurement Test 1

Item No.	Difficulty (p)	Discrimination (r_b)
1	.094	.390
2	.169	.497
3	.136	.475
4	.108	.384
5	.096	.353
6	.153	.384
7	.098	.343
8	.250	.670
9	.267	.538
10	.277	.508
11	.293	.491
12	.295	.460
13	.276	.458
14	.265	.456
15	.210	.451
16	.264	.438
17	.222	.407
18	.205	.398
19	.204	.388
20	.226	.332
21	.220	.326
22	.242	.321
23	.317	.323
24	.318	.348
25	.335	.440
26	.337	.339
27	.345	.612
28	.346	.327
29	.349	.386
30	.353	.375

Table A-3

Item difficulty and discrimination indices for Measurement Test 2

Item No.	Difficulty (p)	Discrimination (r_b)
1	.305	.700
2	.389	.433
3	.299	.403
4	.374	.409
5	.365	.353
6	.386	.349
7	.397	.349
8	.361	.306
9	.398	.396
10	.471	.385
11	.488	.348
12	.445	.333
13	.458	.730
14	.458	.695
15	.458	.637
16	.482	.603
17	.458	.612
18	.458	.611
19	.447	.553
20	.557	.398
21	.537	.398
22	.507	.396
23	.512	.379
24	.585	.369
25	.538	.371
26	.554	.373
27	.553	.354
28	.550	.341
29	.506	.331
30	.542	.307

Table A-4

Item difficulty and discrimination indices for Measurement Test 3

Item No.	Difficulty (p)	Discrimination (r_b)
1	.687	.604
2	.695	.403
3	.677	.540
4	.698	.500
5	.681	.464
6	.686	.474
7	.667	.320
8	.628	.302
9	.749	.610
10	.693	.557
11	.793	.555
12	.795	.581
13	.783	.504
14	.720	.496
15	.721	.495
16	.733	.490
17	.728	.464
18	.719	.457
19	.726	.462
20	.708	.461
21	.708	.457
22	.699	.485
23	.759	.441
24	.754	.438
25	.766	.424
26	.746	.410
27	.791	.373
28	.757	.386
29	.759	.385
30	.788	.377

Table A-5

Item difficulty and discrmination indices for Measurement Test 4

Item No.	Difficulty (p)	Discrimination (r_b)
1	.827	.579
2	.843	.551
3	.811	.550
4	.895	.524
5	.806	.508
6	.857	.487
7	.807	.458
8	.875	.430
9	.850	.405
10	.813	.402
11	.831	.382
12	.884	.367
13	.885	.376
14	.866	.376
15	.890	.367
16	.904	.506
17	.911	.537
18	.921	.565
19	.926	.410
20	.928	.366
21	.942	.385
22	.948	.447
23	.958	.487
24	.963	.560
25	.921	.751
26	.932	.776
27	.937	.693
28	.943	.699
29	.953	.660
30	.958	.710

Appendix B

Item Specifications for Linear Test

Table B-1

Item difficulty and discrimination indices for the linear test

Item No.	Difficulty (p)	Discrimination (r_b)
1	.661	.434
2	.656	.543
3	.659	.490
4	.660	.472
5	.646	.520
6	.646	.477
7	.651	.531
8	.640	.494
9	.634	.534
10	.634	.503
11	.623	.456
12	.610	.518
13	.608	.371
14	.613	.320
15	.607	.516
16	.615	.315
17	.604	.427
18	.602	.538
19	.590	.433
20	.560	.474
21	.557	.448
22	.559	.501
23	.559	.527
24	.549	.496
25	.542	.451
26	.539	.531
27	.542	.490
28	.529	.424
29	.530	.500
30	.514	.448
31	.500	.519
32	.506	.428
33	.449	.520
34	.470	.400
35	.463	.537
36	.439	.466
37	.434	.451
38	.420	.437
39	.419	.482
40	.406	.489