SIMULATION STUDIES OF TWO-STAGE ABILITY TESTING

Nancy E. Betz

and

David J. Weiss

Research Report 74-4

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

October 1974

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

Monte Carlo simulation procedures were used to study the psychometric characteristics of two two-stage adaptive tests and a conventional "peaked" ability test. Results showed that scores yielded by both two-stage tests better reflected the normal distribution of underlying ability. Ability estimates yielded by one of the two stage tests were more reliable and had a slightly higher relationship to underlying ability than did the conventional test scores. One of the two-stage tests yielded an approximately horizontal information function, indicating more constant precision of measurement for individuals at all levels of ability. The conventional test and the second two-stage test yielded information functions peaked at the mean ability level but dropping off at more extreme levels of ability; however, the second two-stage test provided more information than the conventional test at all levels of ability. The findings of the study were interpreted as indicating the potential superiority of two-stage tests in comparison to conventional tests. Several improvements in the construction of two-stage tests are suggested for use in further research.

Contents

SIMULATION STUDIES OF TWO-STAGE ABILITY TESTING

A promising new approach to the measurement of abilities has been made possible by the growth and refinement of time-shared computer facilities. This approach involves varying test item difficulty during the testing procedure according to the estimated ability of the examinee and has been called tailored (Lord, 1970) or adaptive (Weiss & Betz, 1973) testing.

Two-stage testing is one approach to the implementation of adaptive testing procedures. The first stage of a two-stage testing strategy consists of a short "routing" test which is used to obtain a rough initial estimate of the testee's ability. Using this estimate, the testee is then "routed" to a longer second-stage or "measurement" test which consists of items close to his/her estimated ability level. The purpose, then, of two-stage testing is to enable the assignment of each individual to the measurement test most appropriate to his/her ability. Cronbach and Gleser (1965) were the first to suggest the use of two-stage testing procedures. Weiss (1974) 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). Their routing tests were not actually used to assign individuals to measurement tests; rather, measurement tests were embedded within a large sample of items administered to all testees, and the performance of individuals was evaluated on those measurement test items they would have received had routing occurred. Results showed that the measurement tests were more reliable in the sub-groups for which they were intended than were conventional tests measuring a broader range of ability. Predictive validities of the measurement tests, using grade-point average as the criterion, were slightly higher than those of the conventional tests. Their data also showed, however, that 20% of the testees would have been misclassified, or routed into an inappropriate measurement test, on the basis of their routing test score.[1]

A series of "real data" simulation studies of two-stage testing was reported by Cleary, Linn, and Rock (1968 a,b; Linn, Rock, & Cleary, 1969). In these studies, 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 were used to simulate four variations of the two-stage testing strategy.

---

[1]Further information concerning the details of this study and the remaining studies to be discussed may be found in Betz and Weiss (1973), and Weiss and Betz (1973).

Correlations between the artificial 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. In some cases these correlations were higher than the correlations between the parent test and shortened conventional tests using more items than were used in the two-stage tests. The best short conventional test was found to require about 35% more items to achieve the same level of accuracy provided by the two-stage test, and it was concluded that two-stage tests can permit large reductions in the number of items necessary to obtain accurate estimates of ability.

Even more favorable were the findings that the majority of the artificial two-stage tests had higher predictive validities (using scores on the College Entrance Examination Board Tests and the Preliminary Scholastic Aptitude Tests as criteria) than did the conventional tests of the same length. The best two-stage tests had higher validities than longer conventional tests, including the 190-item parent test. These results 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%.

A series of theoretical studies of two-stage testing was reported by Lord (1971c). His analyses were based on the mathematics and assumptions of item characteristic curve theory (Lord & Novick, 1968), including the assumption that the probability of a correct response to an item is a normal ogive function of underlying or latent ability. All items were assumed to be of equal discriminating power, and the items within the routing tests or any one of the measurement tests were assumed to be of equal difficulty.

Lord (1971c) compared the two-stage tests with conventional tests (i.e., tests in which all examinees receive the same items in the same order). However, Lord's conventional tests represented a theoretical ideal in that they were assumed to be perfectly peaked (i.e., all items in a test are of equal difficulty) at the mean ability level of the hypothetical population under study. As in the two-stage tests, all items were also assumed to have equal discriminations. Lord compared the two-stage and conventional tests in terms of information functions, which indicate the relative precision of measurement at various points along the ability continuum. Precision can be defined as the capability of scores based on responses to a set of test items to accurately represent the "true ability" of individuals; the greater the precision at a particular level of ability, the smaller the standard of error of measurement and the confidence interval in estimating true ability at that point.

Lord found that the conventional test provided more precise measurement for ability levels near the group mean, but that the

two-stage procedures provided increasingly better measurement relative to the conventional test with increased divergence from the mean ability level. The finding that the peaked conventional 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" by pyramidal and flexilevel adaptive testing strategies (Lord, 1970, 1971a,b); 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 than did the conventional peaked test.

Figure 1 presents a hypothetical illustration of how the comparative precision or measurement efficiency of conventional and adaptive tests would appear if the values of information at various levels of ability were connected to form a smooth curve. The figure shows that while the conventional peaked test provides superior measurement around the mean ability level, the efficiency of the adaptive tests is more constant across the range of ability and becomes greater than that of the conventional test beyond a given interval containing the mean ability level.

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

An attempt to verify Lord's findings, by routing each individual to that measurement test containing items peaked at median difficulty for him or her, was made in an empirical study of two-stage testing reported by Betz and Weiss (1973). This was the first study to employ computer-administration of test items and computer-controlled routing to the appropriate measurement test within the two-stage paradigm. Each examinee was administered a two-stage test, consisting of a 10-item routing test and one of four 30-item measurement tests, and a 40-item conventional test containing items peaked at the median ability level of the group. The tests were readministered after an interval averaging 5 to 6 weeks in length so that estimates of the test-retest stability could be made.

Results showed that the routing test had as high an internal consistency reliability as did the conventional test, but in contrast to Angoff and Huddleston's (1958) findings, the measurement tests were less reliable than was the conventional test. However, the restriction in ability range caused by the routing procedure would be expected to depress internal consistency reliability. The overall test-retest stability of the two-stage test (.88)

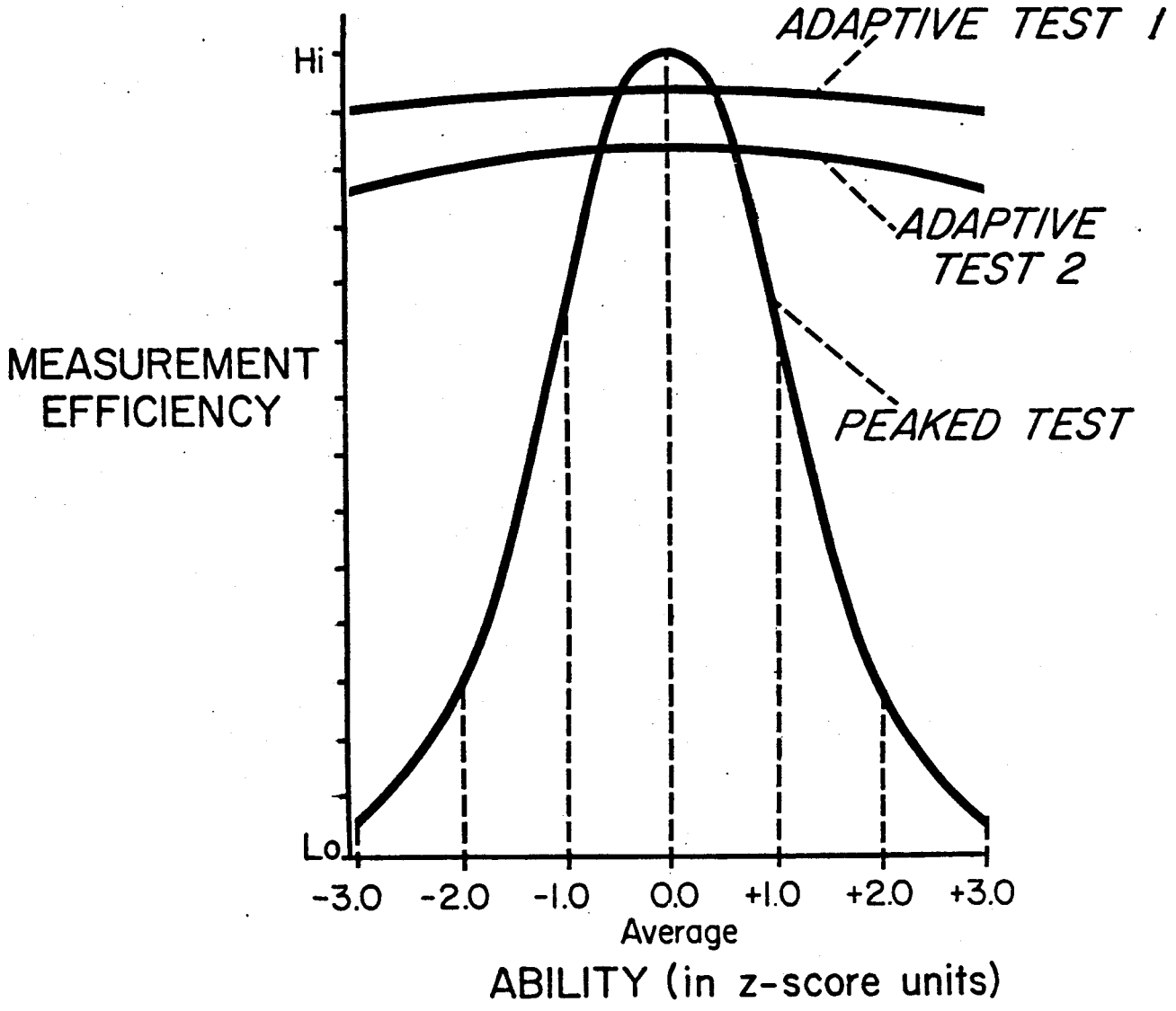


Figure 1. A hypothetical illustration of the comparative measurement efficiency (precision or information) of conventional peaked and adaptive tests.

was as high as that of the conventional test (.89) and was higher (.93) when calculated for only those individuals who had received the same measurement test on the first and second administrations (thus receiving the same opportunity for memory of the previous item responses as was the case with the conventional test).

The routing procedure misclassified only 5% of the testees and was thus an improvement on the 20% rates found in previous studies (Angoff & Huddleston, 1958; Cleary et al., 1968a,b). However, it was also found that the measurement tests were not of optimal difficulty for the groups of individuals assigned to them.

Thus, the studies to date of two-stage testing have shown that it has the potential of providing greater accuracy of measurement and greater predictive validity using fewer items than is possible with conventional tests. However, each of these studies has had limitations which have restricted the generalizability or usefulness of the obtained results. The generalizability of Lord's (1971c) results is limited by the assumption of "ideal" items. Angoff and Huddleston's empirical and Cleary et al.'s "real data" simulation studies are limited by the fact that actual routing did not occur. In the empirical study of Betz and Weiss (1973), the small sample size (N = 214) and the lack of a criterion of "true" ability level prevented the calculation of the relative information or precision of measurement provided by the two-stage and conventional tests.

The present study is, therefore, an attempt to examine the generalizability of the previous findings using Monte Carlo simulation studies of responses to real test items. Monte Carlo studies offer several advantages over other methods of investigating adaptive testing procedures. First, because large numbers of testee "records" can be simulated relatively quickly, it is possible to derive parametric estimates of the characteristics of scores yielded by various testing strategies. These estimates are based on sample sizes sufficiently large to ensure their representativeness. Second, the availability of an ability criterion permits the derivation of information functions and the calculation of their values at points along the hypothetical ability continuum. Third, Monte Carlo simulation studies utilizing two-stage tests composed of items previously administered in empirical studies (e.g., Betz & Weiss, 1973) make it possible to determine whether empirical and simulation studies lead to similar conclusions. Finally, should the results of simulation studies mirror those of the empirical studies, thus validating the simulation model, Monte Carlo methods can then be used to rapidly identify good designs for adaptive testing by providing data concerning the effects of variations in the characteristics of the adaptive testing strategies.

Within the framework of a two-stage adaptive strategy, some of the characteristics which may be varied include: 1) the total number of items given to a single examinee; 2) the number of items in the routing test; 3) the difficulty level of the routing test; 4) the distribution of item difficulties in the routing test; 5) the number of alternative measurement tests available; 6) the cutting points for assigning examinees to measurement tests; 7) the difficulty levels of the measurement tests; and 8) the distributions of item difficulties in the measurement tests. Empirical studies of the promising designs identified in the simulation studies can then be used to evaluate their performance under live testing conditions.

METHOD

Design

The simulation studies were directed at examining the characteristics of the two-stage test and determining whether this testing strategy showed any advantages as compared to conventional ability testing procedures. The simulation studies were designed to permit the investigation of 1) the characteristics of the score distributions yielded by the two-stage and conventional tests in comparison with that of the known ability distribution; 2) the relationships between ability estimates derived from the two-stage strategy and the conventional test; 3) the parallel forms reliability of each test; 4) the relationships between ability estimates and hypothetical underlying ability as specified by the simulation program; and 5) the amount of information or precision provided by each testing strategy at various points along the ability continuum. The first two characteristics examined replicated information obtained in the empirical study (Betz & Weiss, 1973) and were thus considered important to the generalizability of these findings and to the validity of the simulation model. In the simulation study, however, the obtained score distributions could be compared with the known ability distribution.

The third characteristic, parallel forms reliability, had not been studied empirically. Rather, the empirical study examined test-retest stability, or the reliability of the same test over a time interval. However, empirically determined test-retest stability includes as systematic true score variance two sources of error which do not influence simulated test scores. First, the content of an item may contribute error due to specific gaps or emphases in the knowledge of a particular individual. With real subjects, characteristics specific to a particular item are likely to be stable and would thus be reflected in the response on both test and retest. Second, memory of responses made on the first testing may influence the responses of real subjects to the items in the retest. Simulated readministration

of the same test, which was the procedure in this study, is equivalent to the administration of two tests with items whose parameters are identical, and since item content does not influence a simulated item response, the two tests can be thought of as perfectly parallel (Gulliksen, 1950). In the present study, parallel forms reliability was considered to provide a lower-bound estimate of test-retest stability, because errors which act to inflate stability were not present, and also a lower-bound estimate of internal consistency reliability (Guilford, 1954).

The last two areas of interest (relationships with underlying ability and information functions) cannot be studied empirically because underlying ability is not known for real subjects and because the derivation of information functions requires inordinately large sample sizes. Thus, simulation studies make it possible to study important characteristics of the various testing strategies which cannot be studied using other research methods.

Two two-stage tests were studied. Two-stage 1 consisted of the same items that had been administered in the empirical study (Betz & Weiss, 1973). Two-stage 2 was constructed to correct the problems of inappropriate difficulty levels and cutting points that were found in Two-stage 1. The conventional test studied was the same one used in the empirical study (Betz & Weiss, 1973). Each two-stage test was "administered" in conjunction with the conventional test so that the relationships between the resulting score distributions could be found, and all tests were administered twice so that parallel form reliability could be evaluated.

Test administration was simulated for two samples of hypothetical testees. One sample consisted of 10,000 testees whose ability levels were assigned through random sampling from a normally distributed population of ability levels. The second sample consisted of 1,600 testees, 100 at each of 16 discrete ability levels distributed along the ability continuum. This distribution of ability levels, which will be referred to as the "equal-frequency" distribution, was generated for the sole purpose of providing estimates of "information" that were based on equal sample sizes at each selected point on the ability continuum.

Thus, the overall design involved simulated test administration under the following four conditions:

1. Two-stage 1 and the conventional test, each administered twice to 10,000 "examinees" whose ability levels were sampled from a normal distribution of ability levels.

2. Two-stage 2 and the conventional test, each administered twice to 10,000 "examinees" whose ability levels were

sampled (independently of the sample taken in condition 1) from the normal distribution of ability levels.

3. Two-stage 1 and the conventional test, each administered twice to 1,600 "examinees" whose ability levels constituted an "equal-frequency" distribution.

4. Two-stage 2 and the conventional test, each administered twice to the same sample of "examinees" described in condition 3.

Test Construction

Monte Carlo simulation of test administration does not involve the actual administration of test items. Rather, it uses only the input of the relevant item parameters into a formula expressing the relationship between ability level and response to an item with given characteristics. The item parameters selected for input into the simulation program used in this study were those characterizing the items constituting tests constructed for administration to real subjects. The following section, then, describes the manner in which these tests were constructed.

Item Pool

The item pool used to construct the empirical two-stage and conventional tests consisted of five-alternative multiple choice vocabulary items. The items were normed on college students, and normal ogive difficulty ("b") and discrimination ("a") parameters were stored in the computer for each item. Details concerning the development and norming of the item pool are reported by McBride and Weiss (1974).

Two-stage Tests

Each 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 within each subtest (e.g., routing or measurement) were selected to concentrate around a given level of difficulty. While it was not possible to select perfectly peaked subtests given the limitations of a real item pool, the items within each subtest did distribute closely around the desired "b" (item difficulty) value.

Two-stage 1. In the construction of the first empirical two-stage test (Two-stage 1), the difficulty level of the routing test was set to be somewhat easier than the median ability level of the group to account for the probability of chance success on

an item through random guessing (e.g., .2 given 5-alternative responses) as suggested by Lord (1952, 1970). The difficulty levels of the measurement tests were distributed approximately evenly above and below the routing test difficulty.

It was possible to select very highly discriminating items for the routing test because only ten items were required; the measurement tests included items of slightly lower discriminating power. However, highly discriminating items were considered more important in the routing test to ensure the accurate assignment of testees to measurement tests. Table 1 presents the mean item difficulty and discrimination values for the routing test and each measurement test. Both the normal ogive parameters (difficulty, b, and discrimination, a) and traditional item parameters (proportion correct, p, and the biserial correlation with total score, r_b) are presented.

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 the least difficult through the most difficult measurement tests. The lowest score range was the widest since it was expected to include many "chance" scores. Further details on the construction and characteristics of Two-stage 1 may be found in Betz and Weiss (1973).

Two-stage 2. The second two-stage test (Two-stage 2) was constructed to improve on some of the shortcomings of the original two-stage test. First, the routing test was made slightly more difficult (mean b=-.23) since the original routing test (mean b=-.56) had proven too easy for the group as a whole and had created an imbalance in the assignment to measurement tests. Second, the difficulties of the measurement tests were changed in accordance with data concerning the appropriateness of the difficulty levels of the original tests. An examination of Table 1, which summarizes the characteristics of the items of Two-stage 2, shows that in general it was a more difficult test but with a smaller overall spread of item difficulties. Tests 3 and 4, the least difficult measurement tests, were made considerably more difficult than were the corresponding measurement tests in Two-stage 1. And, while the routing test items were as discriminating as those in Two-stage 1, the measurement test items were on the whole somewhat more discriminating. Appendix A gives item reference numbers (see McBride & Weiss, 1974) and difficulty and discrimination values for each item of both two-stage tests.

The routing test score intervals used for assignment to measurement tests in Two-stage 1 were selected on the basis of essentially logical considerations. To formalize and hopefully improve the selection of cutting points for measurement tests in Two-stage 2, the score intervals were determined by calculating a maximum likelihood estimate of ability for each possible routing

Table 1

Summary of item characteristics (norming values) for the two-stage and conventional tests.

Test	Number of Items	Item Difficulty		Item Discrimination	
		Mean "b"	Mean p	Mean "a"	Mean r_b
Two-stage 1					
Routing	10	-.56	.62	.71	.57
Measurement					
1	30	1.81	.24	.47	.42
2	30	.22	.46	.52	.44
3	30	-1.34	.73	.53	.46
4	30	-2.62	.89	.63	.51
Mean		-.49	.58	.55	.47
Two-stage 2					
Routing	10	-.23	.55	.70	.57
Measurement					
1	30	1.73	.23	.53	.46
2	30	.35	.43	.68	.55
3	30	-.71	.64	.61	.52
4	30	-1.60	.80	.68	.55
Mean		-.07	.53	.63	.52
Conventional test	40	-.33	.56	.54	.47

test score (0-10) using the scoring formula described and assigning individuals to that measurement test closest in difficulty to their estimated ability (normal ogive parameter "b" is on the same scale as the ability estimate and thus a direct comparison can be made). The resulting score intervals were 0 through 4, 5 and 6, 7 and 8, and 9-10. Appendix B contains the ability estimates associated with each possible routing score and the resulting measurement test assignment.

Scoring. Two-stage tests cannot be scored using a simple number-correct score since examinees take different measurement tests having different difficulty levels. The method of scoring two-stage tests suggested by Lord (1971c) takes both the number correct and the difficulty level of the items into account. It consists of obtaining two maximum likelihood estimates of ability (θ), one from the routing test ($\hat{\theta}_1$) and one from the appropriate measurement test ($\hat{\theta}_2$). These two estimates are then averaged after weighting them inversely according to their estimated variances.

The formula used by Lord to obtain estimates of θ from the routing and measurement tests was as follows:

$$\theta = \frac{1}{a}\Phi^{-1}\left[\frac{(x/m) - c}{1 - c}\right] + b \qquad (1)$$

where a is the normal ogive discrimination value of the items;

x is the number correct;

m is the total number of items administered in that subtest;

c is the chance-score level;

b is the normal ogive difficulty level of the items in the subtest;

Φ^{-1} (the inverse of Φ) is the relative deviate corresponding to a given normal curve area.

In the present study, equation 1 was modified slightly to account for the fact that the items in any given subtest were not all of equal discrimination and difficulty. The formula used was as follows:

$$\hat{\theta} = \frac{1}{\bar{a}}\Phi^{-1}\left[\frac{(x/m) - \bar{c}}{1 - \bar{c}}\right] + \bar{b} \qquad (2)$$

where $\bar{a}$ represents the mean discrimination value, and

$\bar{b}$ the mean difficulty of the items in that subtest.

The value of c was always .2 since the items had five alternative responses. 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=m-.5=9.5, and when x was less than or equal to cm, it was replaced by x=cm+.5=2.5.

The two ability estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ for the routing and measurement test respectively as computed from equation 2 were combined into a total-test ability estimate by averaging the two after weighting each by the number of items (10 or 30) on which it was based. This method of weighting was used instead of the variance weights used by Lord (1971c) since the latter method was found to have some disadvantageous characteristics (see Betz & Weiss, 1973). The composite ability estimate, then, was defined by the following equation:

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

Scores determined in this way can be interpreted similarly to standard normal deviates, i.e., they have a mean of 0 and a variance of 1.

Conventional Test

The conventional test consisted of 40 items. As in construction of the two-stage tests, the use of a real item pool did not permit the construction of a perfectly peaked or equidiscriminating test as had been studied by Lord (1971c). Item difficulties were concentrated around a "b" value of -.33 (again somewhat easier than the median ability level of the group). While the range of difficulties was large for a peaked test, it was small in relation to the range of difficulties covered by all four of the second-stage measurement tests used in either Two-stage 1 or Two-stage 2. Table 1 also summarizes the characteristics of the 40-item conventional test. Appendix A gives difficulty and discrimination values for each item of the conventional test. Additional details on the construction of this test may be found in Betz and Weiss (1973, p.15). Number correct was used as the score on the conventional test.

Simulation of Test Responses

The Simulation Model

Development of the simulation procedure was based on the assumptions and mathematics of item characteristic curve theory (Lord & Novick, 1968). Using the mathematical model suggested by Lord (1970, 1971c), the probability of a correct response to an item was assumed to be a generalized normal ogive function

of the examinee's ability and was determined through the solution of the following equation:

$$P_i(\theta) = c_i + (1 - c_i)\,\Phi\left[a_i(\theta - b_i)\right] \qquad (4)$$

In this formula, $P_i(\theta)$ is the probability that an examinee with ability θ will respond correctly to item i. The a_i, b_i, and c_i are the normal ogive parameters of item i, where a_i represents the discriminating power of the item, b_i represents the difficulty level of the item, c_i is the guessing parameter, or the probability that the item can be answered correctly through random guessing, and $\Phi[x]$, the cumulative normal distribution function, represents the normal distribution cumulative proportion up to the relative deviate x.

In the solution of equation 4, c_i was set at .2 since the item pool was based on five-alternative multiple-choice items. Difficulty and discrimination parameters were those associated with each item administered (see Appendix A), and ability level was specified as described below.

Procedure

Appendix C describes the computer program which simulated test administration and calculated test scores. The data yielded by the program consisted of ability level and four ability estimates (two each from the two-stage and conventional tests) for each hypothetical subject. Each "run" of the program provided data for 100 hypothetical individuals; following each "run," the Pearson product-moment correlations among test scores and between test scores and underlying ability were calculated for that group of 100 "testees."

Underlying ability level was specified in two ways. To obtain a subject population with a normal distribution of abilities, a pseudo-random number generator yielding a normally distributed set of numbers with a mean of 0 and a variance of 1 was used to assign an ability level to each of 10,000 hypothetical individuals. To obtain the "equal-frequency" distribution of ability, each of 16 ability levels between $\theta=-3.2$ and $\theta=+3.2$ were assigned to 100 individuals. The 16 ability levels used are shown in Table 7.

Once ability level had been specified, item "administration" was begun. The parameters of the particular item to be administered were entered, along with the ability level, into equation 4 to calculate the probability ($P_i(\theta)$) of a correct response to that item. Following the calculation of $P_i(\theta)$, a random number P from a rectangular distribution of real numbers between 0 and 1 was generated. If $P < P_i(\theta)$, the item was scored "1" (correct), and if $P > P_i(\theta)$ the item was scored

"0" (incorrect). The item response, 1 or 0, was then stored in the computer for use in scoring the test.

In the conventional test, the items were administered in the order shown in Appendix Table A-3. In the administration of the two-stage tests, the routing test score (number correct of the first ten items administered) was calculated and the next thirty items administered were those constituting the appropriate measurement test, using the routing rules described previously for each of the two-stage tests.

Analysis of Data

The basic set of data to be analyzed consisted of ability level, two scores on Two-stage 1, and two scores from the conventional test for each of 10,000 "testees." For the second group of 10,000 "testees," the data consisted of ability level, two Two-stage 2 scores, and two conventional test scores. Analysis of the former data set was designed to replicate the analyses of the live-testing study reported by Betz & Weiss (1973) using the same two-stage test (Two-stage 1) and the same conventional test.

While it was assumed that samples of 10,000 ability levels generated from a normally distributed population would be normally distributed, the characteristics of the two resulting distributions of ability were analyzed to determine whether or not this assumption was valid. For each distribution of 10,000 ability levels, the mean, variance, and the degrees of skewness and kurtosis were calculated. The degrees of skewness and kurtosis were tested for the significance of their departure from normality (McNemar, 1969, pp. 25-29 and 87-88). Both distributions of ability were found to be normal. The means were 0.0, and the variances were 1.0. The degree of skewness was .010 for both distributions (as compared to the standard error of .025 given an N of 10,000). The degree of kurtosis was -.003 for the first distribution and -.04 for the second distribution (as compared to a standard error of .05).

A second set of data consisted of ability level and the same two sets of four scores as described above for 1600 "testees," 100 at each of 16 ability levels. This data was used only in the calculation of values of the information functions at each of the 16 ability levels, while the data obtained from the two groups of 10,000 "testees" were used in all analyses to be described.

Characteristics of Score Distributions

Analyses of the characteristics of the score distributions were done separately for the two administrations (test and retest) of each test. For each distribution of 10,000 scores, the score mean, standard deviation, and the degrees of skewness

and kurtosis were calculated; the degrees of skewness and kurtosis were tested for the significance of their departure from normality (McNemar, 1969, pp. 25-28 and 87-88).

Parallel Forms Reliability

Pearson product-moment correlation coefficients were calculated to express the degree of relationship between scores obtained from the two administrations of each testing strategy for each group of 100 individuals. Thus, there were 100 reliability coefficients obtained from 100 samples from a hypothetical population having a normal distribution of underlying ability. The sampling distribution of these coefficients was used to make inferences to the expected value of the population value ρ and to construct confidence intervals within which ρ could be expected to fall in 95% of such sampling experiments. The expected value of ρ was taken as the mean of the distribution of 100 r values, and 95% confidence intervals were obtained by adding and subtracting from the expected value a value equal to two standard deviations of the obtained sampling distribution. Fisher's z-transformation was applied to each sample value of r, and the sampling distribution of values of Z_r was also obtained. Confidence intervals were then calculated using ± 2 standard deviations of this distribution, and the resulting values were transformed into their corresponding values of r.

Interrelationships among Test Scores and between Scores and Underlying Ability

Product-moment correlation coefficients were calculated between scores on Two-stage 1 and the conventional test and between scores on Two-stage 2 and the conventional test; the total score distributions of 10,000 scores were used in this analysis. In addition, eta coefficients were calculated for each total score distribution regressed on the other one, again using all 10,000 scores obtained from each testing strategy; tests of curvilinearity were made to determine if there were non-linear relationships between score distributions.

Similar analysis using both Pearson product-moment and eta correlation coefficients was done to determine the nature and degree of the relationship between Two-stage 1, Two-stage 2, and conventional test scores and ability level for all 10,000 subjects. Thus, the values of r obtained using an N of 10,000 provided one estimate of the expected value of ρ in the population. The characteristics of the sampling distributions of the 100 product-moment coefficients and Z-transformed r's calculated on each group of 100 testees were also calculated and used to obtain expected values and confidence intervals for the ρ values.

Information Functions

The information function is used to compare two or more strategies of testing in terms of the amount of information (or relative degree of precision of measurement) provided at different levels on the ability continuum. The value of information at each level of underlying ability was calculated using the formula suggested by Birnbaum (1968):

$$I_x(\theta) = \left[\frac{\frac{\partial}{\partial \theta} \varepsilon(x|\theta)}{\sigma_{x|\theta}} \right]^2 \qquad (5)$$

where $I_x(\theta)$ indicates the amount of information provided by test X, scored in some specific way, at a given level of underlying ability θ. The numerator in equation 5 is the slope of the regression of observed test scores on underlying ability (calculated by solving the equation for the first derivative for that value of θ), and the denominator is the standard deviation of test scores obtained by testees with ability θ. This ratio is then squared to obtain $I_x(\theta)$.

The numerator of equation 5 represents the capability of test scores to differentiate among examinees of different levels of underlying ability. For example, given examinees at two levels of ability θ_1 and θ_2 and expected test score values x_1 and x_2, the magnitude of the slope

$$\frac{x_2 - x_1}{\theta_2 - \theta_1} \qquad (6)$$

indicates the degree to which the test discriminates these two ability levels. The denominator of equation 5 is the precision of measurement at a particular level of ability. The square root of $I_x(\theta)$ is inversely related to the confidence interval for estimating underlying ability from observed score (Green, 1970). Thus, a low value of $I_x(\theta)$ indicates a larger confidence interval and a larger standard error of measurement at a particular level of ability, and the higher the value of $I_x(\theta)$, the narrower the confidence interval and the smaller the error of measurement. Information values are not meaningful in any absolute sense because they are dependent on the scale used to measure θ and also on the scoring formula used to determine x, but information values calculated from two or more strategies assuming the same θ scale can be directly compared, with larger

values indicating more precise measurement.

The relative amount of information provided by the two-stage and conventional tests was calculated for both the normally distributed and "equal-frequency" distributions of ability. The regression equation relating test score (the dependent variable) to generated ability (the independent variable) was calculated from the normal distribution data using a least squares curve-fitting computer program. The third degree or cubic polynomial equation generated was used since higher degree polynomial equations did not significantly reduce the standard error of estimate of the dependent variable (i.e., test score). The slope function for each test was obtained by taking the first derivative of the third degree polynomial equation describing its regression on generated ability.

The normal ability distribution was divided into 33 intervals between $\theta = -3.3$ to $\theta = +3.3$. Each interval had a width of .2, and the midpoint of the interval was used to calculate the slope of the function at that level of ability. Thus, the lowest ability interval was $\theta = -3.3$ to $\theta = -3.1$, and $\theta = -3.2$ was taken as its midpoint. For each interval, the variance of the test scores of individuals whose hypothetical ability level fell into that interval was calculated.

When the normal distribution of ability was used, however, the number of individuals within each interval differed at all points along the ability continuum. That is, since interval length was constant, large numbers of individuals fell into the intervals in the middle of the continuum, while the ability intervals at or near the extremes had considerably fewer individuals. Thus, information values for extreme ability levels were less stable than those nearer the middle because the score variance was more influenced by chance similarities or differences among scores determined for individuals of approximately the same ability.

As a result, the "equal-frequency" distribution of ability was used to obtain information values of equal stability or reliability at all points along the ability continuum. The slope value used was that generated from the normal distribution and was computed at each of the 16 ability levels indicated in Table 7. Thus, the numerator of the information equation was the squared slope at each ability level, and the denominator was the variance of the 100 scores generated at that level.

Since each test was administered twice to each sample of "testees," there were two sets of information values for each test. These values were averaged to obtain an overall index of information at each ability level for each test. Finally, the mean and standard deviation of each set of 33 information values (obtained from the normal distribution of ability) and 16 values (obtained from the "equal-frequency" distribution)

were calculated. The mean of each set was interpreted as an index of the general level of information provided by each test, while the standard deviation was considered to provide an indication of the constancy of information provided for the entire range of ability levels sampled.

RESULTS

Score Distributions

Table 2 presents data describing the distributions of scores obtained from the two-stage tests and the conventional test. Data are presented for both administrations (test and retest) of each test. Since the data derived from administration of the conventional test with Two-stage 2 were identical to those of the test when administered with Two-stage 1, only the latter set of results is presented. These values can be considered as representative indicators of the characteristics of the conventional test used in this series of studies.

Two-stage 2, the improved two-stage test, resulted in a distribution of scores which better reflected the underlying distribution of ability (normal with mean 0 and variance 1) than did Two-stage 1. The mean score on Two-stage 2 was essentially 0, and the standard deviations (1.06 and 1.05) were closer to 1.0 than those of Two-stage 1 (1.24 and 1.22). The skewness of the Two-stage 2 score distribution did not show a significant departure from normality, while the distribution of Two-stage 1 scores was significantly skewed in the negative direction. While the Two-stage 2 distribution was significantly more platykurtic (flat) than a normal distribution, the degrees of kurtosis (-.20 and -.23) were less than those of Two-stage 1 (-.42 and -.49).

Both two-stage tests showed less skewness than did the conventional test, in which scores were significantly negatively skewed (-.25 and .-23). The conventional test score distribution was also platykurtic, to about the same degree as that of Two-stage 1 and to a greater degree than that shown by Two-stage 2. Thus, the score distribution yielded by Two-stage 2 better reflected the underlying normal distribution of ability than did the conventional test.

The score distributions yielded by Two-stage 1 and the conventional test in the empirical study (Betz & Weiss, 1973) were not skewed; both distributions, however, tended toward platykurtosis, and this tendency was statistically significant in the conventional test scores in the empirical study.

Parallel Forms Reliability

Table 3 presents the characteristics of the sampling distribution of parallel forms reliability coefficients. Again, the results from the conventional test were identical for the

Table 2

Descriptive data for two-stage and conventional test score distributions under the assumption of a normal distribution of ability (N = 10,000)

Test	Mean	Standard Deviation	Skew	Kurtosis
Two-stage 1				
Time 1	.062	1.24	-.08*	-.42*
Time 2	.087	1.22	-.09*	-.49*
Two-stage 2				
Time 1	-.004	1.06	-.04	-.20*
Time 2	-.004	1.05	-.04	-.23*
Conventional				
Time 1	25.9	6.48	-.25*	-.46*
Time 2	25.9	6.43	-.23*	-.53*

*Significant at $p<.01$

Table 3

Characteristics of sampling distribution of parallel forms reliability coefficients using 100 random samples of 100 "testees"

	Mean	S. D.	Range		95% Confidence Interval (± 2 S. D.'s)			
			Maximum	Minimum	Upper	Lower	Upper*	Lower*
Two-stage 1	.76	.045	.87	.63	.85	.67	.84	.66
Two-stage 2	.83	.034	.89	.72	.90	.76	.89	.75
Conventional	.80	.038	.88	.65	.87	.72	.86	.72

*Based on a sampling distribution of Z-transformed r's.

administrations with Two-stage 1 and Two-stage 2, so only one set of data is presented for the conventional test.

Table 3 shows that Two-stage 2 was more reliable ($\bar{r}$=.83) than either the conventional test ($\bar{r}$=.80) or Two-stage 1 ($\bar{r}$=.76). Further, the range and variability of the distribution of reliability coefficients was smallest for Two-stage 2, indicating more consistency in reliability estimates determined from the 100 samples. The obtained confidence intervals indicate that the reliability of Two-stage 1 is probably between .66 and .85, with an expected value of .76. The reliability of Two-stage 2 is probably between .75 and .90 (expected value .83), and that of the conventional test probably falls in the interval between .72 and .87, with an expected value of .80.

Relationships between Two-stage and Conventional Test Scores

Table 4 presents the linear (product-moment) correlations and eta coefficients describing the relationship between scores on each two-stage test and conventional test scores. All of the coefficients were significantly different from zero ($p<.001$) and indicate a high and predominantly linear relationship between scores obtained from the two methods of testing. Although two of the eta coefficients indicated a significant degree of curvilinearity, the absolute increase in the degree of relationship with curvilinearity taken into account was very small and not practically significant; with a sample size of 10,000 very small curvilinear trends may attain statistical significance.

Two-stage 2 showed a higher degree of linear relationship (r=.82) with the conventional test scores than did the original two-stage test (r=.78 or .79) and thus accounted for an additional 6% (67% versus 61%) of the variance in the conventional test scores. These values may be compared with those obtained in the live-testing study of Two-stage 1 and the same conventional test (Betz & Weiss, 1973), where the linear relationships between the tests were r=.80 and r=.84 on test and retest, respectively, thus accounting for 64% and 70% of the variance. These values compare quite closely to the values obtained in the present study, and, similarly, there was no evidence for important curvilinear trends in the empirical data.

Relationships between Test Scores and Ability

Table 5 presents the degree of linear and curvilinear relationship between test scores and generated ability level when calculated using all 10,000 scores obtained from each testing strategy. All of the coefficients were significant at $p<.001$ and, again, the relationships were high and predominantly linear. Examination of the bivariate scatter plots did not show clear curvilinear trends, and the eta coefficients do not add importantly to the degree of linear relationship found.

Table 4

Regression analysis of relationships between two-stage scores and conventional scores (N=10,000)

Test and Index of Relationship	Time 1	Time 2
Two-stage 1 and conventional		
Product-moment correlation	.79	.78
Regression of two-stage scores on conventional scores (eta)	.79	.78
Regression of conventional scores on two-stage scores (eta)	.79	.78*
Two-stage 2 and conventional		
Product-moment correlation	.82	.82
Regression of two-stage scores on conventional scores (eta)	.82*	.82
Regression of conventional scores on two-stage scores (eta)	.82	.82

*Degree of curvilinearity significant at $p < .001$.

Table 5

Regression analysis of relationships between test scores and ability (N=10,000)

Test and Index of Relationship	Time 1	Time 2
Two-stage 1		
Product-moment correlation	.87	.87
Regression of two-stage scores on ability (eta)	.87	.87
Regression of ability on two-stage scores (eta)	.87	.87*
Two-stage 2		
Product-moment correlation	.91	.91
Regression of two-stage scores on ability (eta)	.91	.91
Regression of ability on two-stage scores (eta)	.91	.91
Conventional		
Product-moment correlation	.90	.90
Regression of conventional test scores on ability (eta)	.90*	.90*
Regression of ability on conventional test scores (eta)	.90	.90*

*Curvilinearity statistically significant at $p<.005$.

Table 6

Characteristics of sampling distributions of relationship of test scores to ability: product-moment correlations calculated on 100 random samples of 100 hypothetical individuals

Variables	Mean	S. D.	Range		95% Confidence Interval (±2 S. D.'s)			
			Maximum	Minimum	Upper	Lower	Upper*	Lower*
Two-stage 1-- ability (Time 1)	.87	.027	.93	.77	.92	.82	.92	.81
Two-stage 1-- ability (Time 2)	.87	.025	.93	.82	.92	.82	.92	.82
Two-stage 2-- ability (Time 1)	.91	.024	.95	.82	.95	.86	.94	.86
Two-stage 2-- ability (Time 2)	.91	.016	.95	.86	.94	.88	.94	.87
Conventional -- ability (Time 1)	.89	.020	.93	.81	.93	.85	.93	.85
Conventional-- ability (Time 2)	.89	.019	.93	.85	.93	.85	.93	.85

*Based on sampling distribution of Z-transformed r's.

Two-stage 2 showed the highest relationship to underlying ability (r=.91), followed by the conventional test (r=.90) and Two-stage 1 (r=.87). Thus, underlying ability level accounted for approximately 83% of the variance in Two-stage 2 scores, 81% of the variance in conventional test scores, and 76% of the variance in scores on Two-stage 1.

Table 6 presents the characteristics of the sampling distributions of the obtained and Z-transformed product-moment coefficients calculated on 100 groups of 100 testees. A comparison of the mean values shown in Table 6 with the values in Table 5, calculated only once for 10,000 testees, shows that they are identical except for the conventional test, where the mean value of 100 coefficients is .89 (Table 6) and the value for all 10,000 testees (Table 5) is .90.

Examination of the confidence intervals within which the true population correlation (ρ) may be expected to fall shows that the two methods of calculation, using the obtained distribution of r or the distribution of Z-transformed r's, yield very similar results. The Z-transformed coefficients yield an interval of between .81 and .92 for the true relationship between Two-stage 1 scores and generated ability, .86 to .94 for Two-stage 2 scores and generated ability, and .85 to .93 for the relationship between scores on the conventional test and underlying ability.

Information Functions

Equal-frequency distribution. Table 7 presents the values of the information function ($I_x(\theta)$) for the two-stage and conventional tests at each of sixteen ability levels. The value at each level represents the average of the values obtained from the two administrations of each test; separate values for the first and second administrations may be found in Appendix Table D-1. These values may be compared directly among tests and are equally reliable for each ability level. Table 7 also presents the mean and standard deviation of the 16 values obtained for each test. The data contained in Table 7 are summarized in graphic form in Figure 2; the point values have been connected and the curves visually smoothed to convey the shape of the information functions for the three tests. (The unsmoothed information functions for the three tests are contained in Appendix E).

The shape of the information curve for the conventional test, as shown in Figure 2, is very similar to that found in Lord's (1971c) theoretical study; that is, the information values are highest at the center of the ability distribution and drop off sharply at the extremes. Both Lord's results, using "ideal" items, and the results indicated here, using a set of items with parameters that are typical of those occurring in empirical test construction and which did not permit the con-

Table 7

Values of the information function ($I_x(\theta)$) for two-stage and conventional tests at points along the continuum of underlying ability (equal-frequency distribution of ability, N=200 at each level)

Level of Ability (θ)	Two-stage 1	Two-stage 2	Conventional
3.2	2.51	2.03	.85
3.0	2.43	1.73	.02
2.5	3.53	3.47	1.12
2.0	4.09	3.98	3.29
1.5	3.59	4.38	4.38
1.0	2.95	4.96	4.42
.5	3.12	5.72	5.30
.1	2.67	6.22	4.46
-.1	2.86	4.75	4.38
-.5	3.60	4.91	4.25
-1.0	3.86	5.22	3.53
-1.5	3.66	4.76	3.01
-2.0	3.19	2.58	2.50
-2.5	2.43	2.94	1.32
-3.0	2.41	1.13	.39
-3.2	2.10	.79	.15
Mean	3.06	3.72	2.71
S.D.	.61	1.68	1.81

struction of a perfectly peaked conventional test, show that a conventional test offers greatest precision of measurement for individuals near the median ability level of the group and decreasing precision with divergence of an individual's ability from the median level.

Figure 2 shows that Two-stage 1 provided more constant information across ability levels than did either Two-stage 2 or the conventional test; it provided less information around the median ability level but more information at the extremes. The results for Two-stage 1 were similar to those found by Lord (1971c) in his theoretical study of two-stage tests. However, the results for the improved two-stage test were quite different from those obtained in Lord's theoretical studies. The information curve for Two-stage 2 was more similar in shape to that of the conventional test, showing greatest precision in the center of ability distribution and a loss in precision at the extremes. However, at every ability level, its information values were higher than those of the conventional test.

The overall level and shape of the information functions shown in Figure 2 are also reflected by the means and standard deviations of the information values for each test, as shown in Table 7. The average value for Two-stage 2 was 3.72, higher than that for Two-stage 1 (3.06) and the conventional test (2.71). The tendency of Two-stage 1 to yield a horizontal information function rather than a peaked one, indicating more even or constant precision of measurement, is reflected by the small standard deviation of information values (.61) as compared to that of Two-stage 2 (1.68) and the conventional test (1.81).

One way to interpret information values is in terms of the relative numbers of items necessary to achieve equivalent precision of measurement for a given individual. For example, if for a specified level of ability, information for Test A is twice as great as the value of information for Test B, it indicates that Test B would require twice as many items as would Test A to achieve the same level of precision of measurement. Thus, the values shown in Table 7 indicate that at θ=2.5, the conventional test would require nearly three times as many items to achieve the same level of precision as provided by Two-stage 2 for individuals of that ability. At θ =.1, the conventional test would require 39% more items, at θ=-1.0 it would require 47% more, and at θ =-2.5, the conventional test would require over twice the number of items.

Examination of the points at which the three curves shown in Figure 2 intersect indicates comparative information or precision for ranges of ability. Two-stage 1 and Two-stage 2 intersect at about θ=-2.0 and θ =+2.0; Two-stage 2 was superior within this range, and Two-stage 1 was superior beyond it. Two-stage 1 was superior to the conventional test when $\theta > +1.5$ and $\theta < -1.0$, and Two-stage 2 was superior to the conventional test at all levels of ability. Thus, of the three tests,

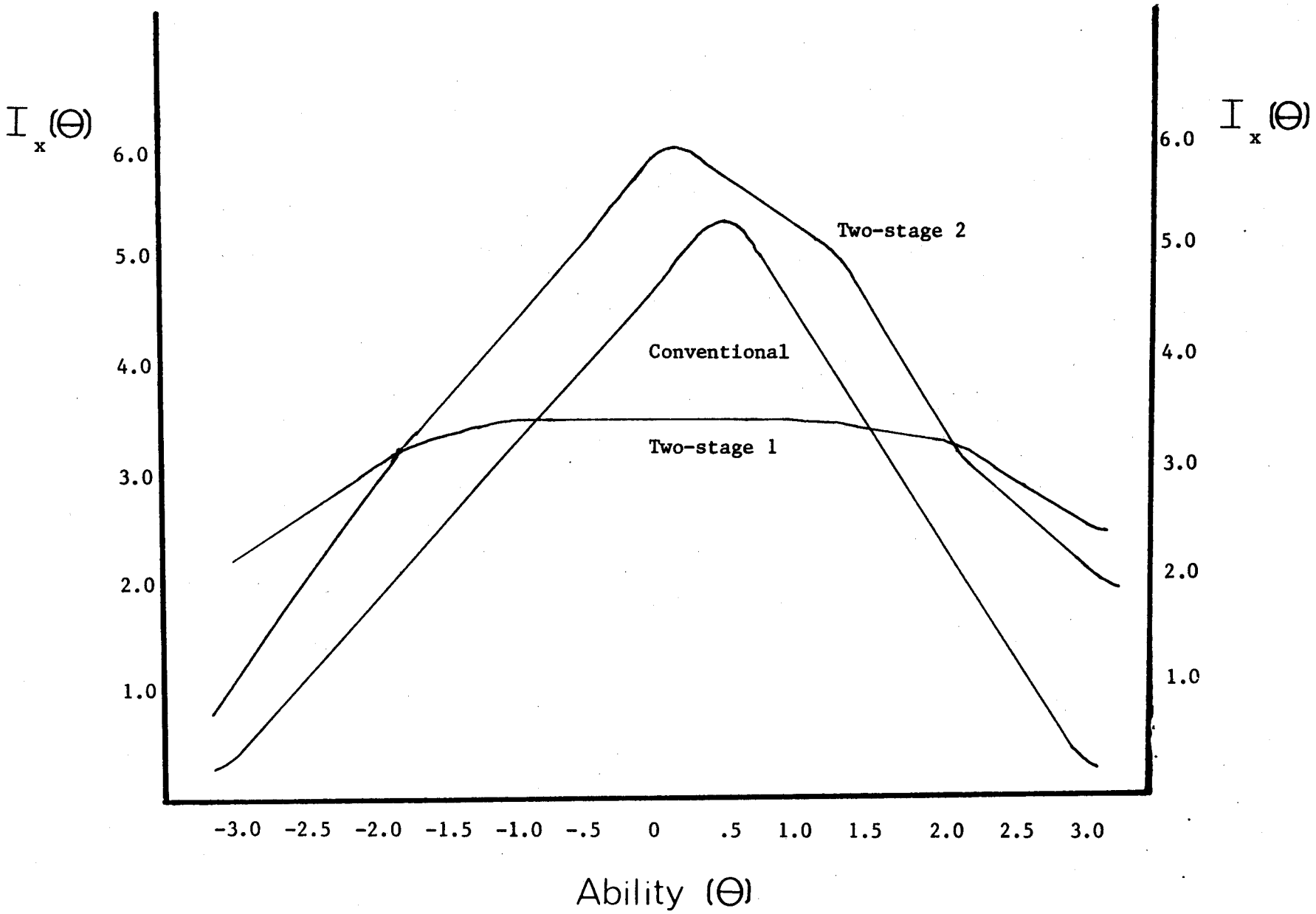


Figure 2. Smoothed information functions of two-stage and conventional tests using "equal-frequency" distribution of underlying ability.

Two-stage 2 provided most precise measurement (least amount of error) for testees whose abilities were between -2.0 and +2.0 standard deviations from the population average, and Two-stage 1 provided more accurate measurement for testees whose ability was beyond this range. These data indicate that at least one of the two-stage tests provided more accurate measurement than the conventional test at all levels of ability.

Normal ability distribution. Table 8 presents the values of $I_x(\theta)$ provided by the two-stage and conventional tests under the assumption of a normal distribution of ability; again, these values represent the average of the values obtained from the two administrations of each test (separate values of the first and second administrations may be found in Appendix Table D-2). Table 8 also indicates the total number of "testees" upon which each value of $I_x(\theta)$ is based. For example, only two "testees" were assigned an ability level between $\theta = 3.1$ and $\theta = 3.3$ in the Two-stage 1 administration. Thus, with two administrations of the test, the $I_x(\theta)$ value is based on a total of four test administrations. Obviously, the $I_x(\theta)$ values based on N's of 4, 14, or 30 at extreme ability levels cannot be considered to be as representative of the true information value at that ability level as may the $I_x(\theta)$ values for abilities near the mean which were based on N's of 1500 or 1600. Again, the mean and standard deviation of the values for each test are presented.

The results indicated in Table 8 are summarized graphically in Figure 3, which shows the smoothed information curves for the three tests. Appendix E (Figure E-2) shows the raw curves for the normal distribution data.

Given the differences in the reliability of the values determined from the normal and "equal-frequency" distributions of ability, the results are remarkably similar. As shown in Figure 3, Two-stage 1 again displays very constant information along the ability continuum, while Two-stage 2 and the conventional test show high levels of precision around the median ability level but losses of precision at the extremes. Two-stage 2 again, however, provides more information at all levels of ability than does the conventional test.

The means and standard deviations of the information values shown in Table 8 indicate that Two-stage 2 provided the highest overall level of information (3.89), but that Two-stage 1 provided almost as high an average value (3.59). However, Two-stage 1 had substantially less variability in the distribution of obtained values (.96 as compared to 1.36 for Two-stage 2). The conventional test provided the least amount of information overall (2.86) and its values were the most variable (1.57), indicating least tendency toward constant precision across the ability continuum.

Table 8

Value of the information function for two-stage and conventional tests within intervals of the continuum of underlying ability under the assumption of a normal distrubution of ability (values are the average of two administrations and are based on the indicated total numbers of hypothetical individuals)

Interval of Ability (θ)	Two-stage 1		Two-stage 2		Conventional	
	N	$I_x(\theta)$	N	$I_x(\theta)$	N	$I_x(\theta)$
3.1 to 3.3	4	4.51	8	.82	4	.58
2.9 to 3.1	14	8.35	14	1.52	14	.23
2.7 to 2.9	30	4.28	24	2.69	30	.06
2.5 to 2.7	52	2.29	70	3.92	52	.42
2.3 to 2.5	100	3.53	82	3.61	100	1.66
2.1 to 2.3	168	3.90	160	3.72	168	1.93
1.9 to 2.1	192	3.14	198	3.75	192	3.48
1.7 to 1.9	318	3.97	310	3.71	318	3.73
1.5 to 1.7	452	3.72	420	3.95	452	3.55
1.3 to 1.5	596	3.41	590	4.70	596	4.26
1.1 to 1.3	742	3.77	732	5.19	742	4.20
0.9 to 1.1	1042	3.51	1016	4.82	1042	4.77
0.7 to 0.9	1088	3.31	1168	4.81	1088	4.36
0.5 to 0.7	1334	3.09	1272	4.20	1334	4.42
0.3 to 0.5	1496	2.91	1488	4.75	1496	4.87
0.1 to 0.3	1442	2.80	1542	5.05	1442	4.55
-.1 to 0.1	1690	3.07	1662	5.31	1690	4.78
-.3 to -.1	1548	3.00	1640	5.61	1548	4.44
-.5 to -.3	1550	3.08	1442	5.51	1550	4.25
-.7 to -.5	1264	3.57	1286	5.65	1264	3.96
-.9 to -.7	1156	3.20	1188	5.36	1156	3.78
-1.1 to -.9	948	3.22	926	5.27	948	3.48
-1.3 to -1.1	652	3.41	782	4.77	652	3.56
-1.5 to -1.3	660	3.49	598	4.25	660	3.36
-1.7 to -1.5	470	3.57	424	4.39	470	3.10
-1.9 to -1.7	350	3.42	302	4.01	350	2.74
-2.1 to -1.9	208	3.28	202	3.32	208	2.54
-2.3 to -2.1	144	3.90	176	2.79	144	1.98
-2.5 to -2.3	82	2.97	92	1.90	82	1.48
-2.7 to -2.5	56	3.48	46	2.73	56	2.02
-2.9 to -2.7	40	3.91	28	1.35	40	.80
-3.1 to -2.9	16	3.74	10	1.20	16	.84
-3.3 to -3.1	12	3.64	8	3.81	12	.15
Mean		3.59		3.89		2.86
S.D.		.96		1.36		1.57

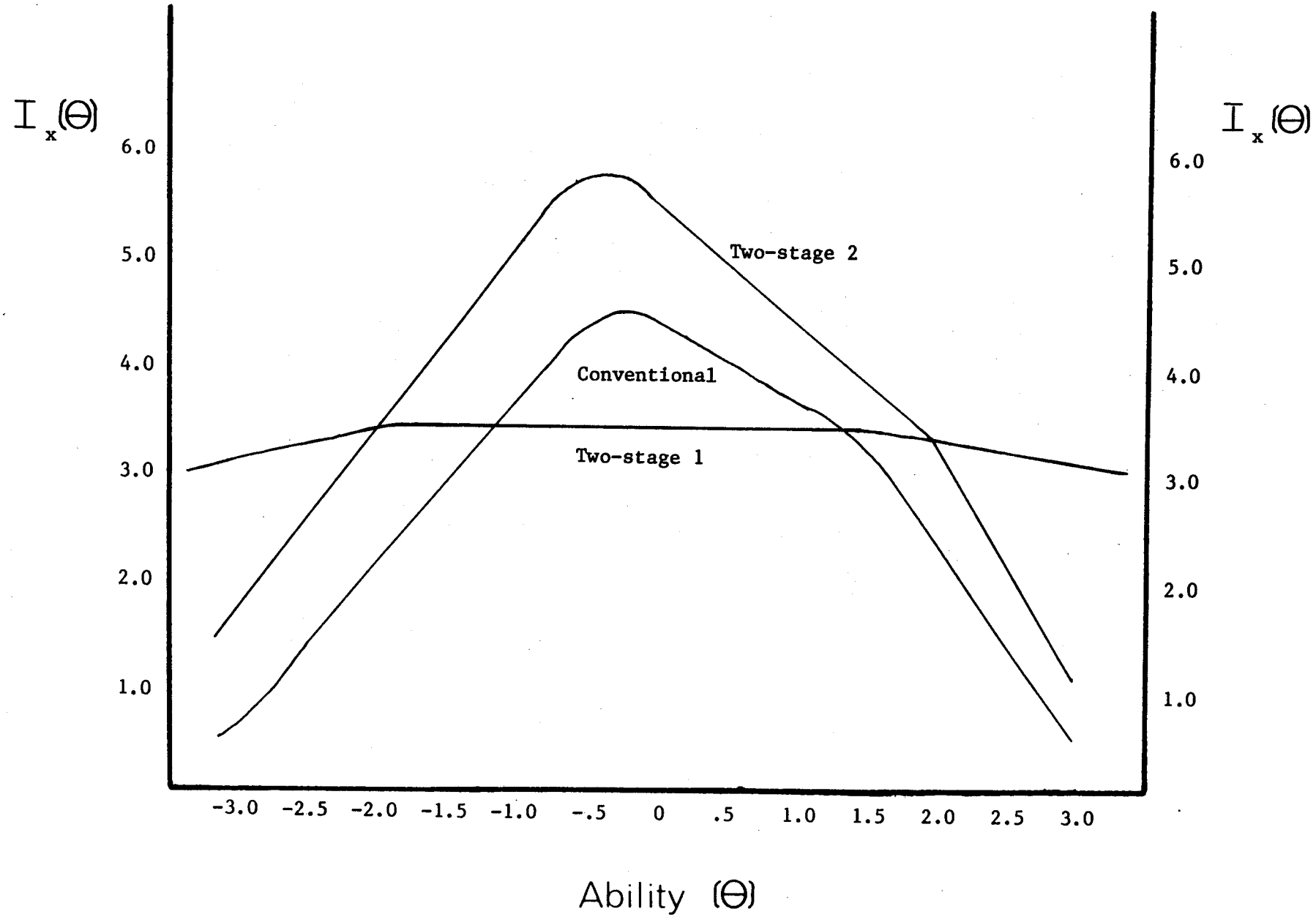


Figure 3. Smoothed information functions of two-stage and conventional tests using normal distribution of underlying ability.

The intersections of the information curves in Figure 3 again show that Two-stage 2 provided more information than Two-stage 1 within the interval $\theta = -2.0$ to $\theta = +2.0$, while Two-stage 1 was superior beyond this interval. Two-stage 1 was superior to the conventional test at ability levels greater than $\theta = +1.5$ and less than $\theta = -1.5$, while Two-stage 2 was superior to the conventional test at essentially all levels of ability.

In general, the information functions derived from both the normal and "equal-frequency" distributions of ability show that Two-stage 2 provided most precision for ability levels within two standard deviations of the mean ability level, while Two-stage 1 provided most precision outside that range.

CONCLUSIONS

Both two-stage tests yielded score distributions which better reflected the normal distribution of ability than did the conventional test. However, the improved two-stage test (Two-stage 2) was superior in this regard to both the original two-stage test (Two-stage 1) and the conventional test. All score distributions showed a significant degree of platykurtosis and were thus flatter or more rectangular than the normal distribution. This may be explained by the fact that the two-stage test is designed to "spread" people out by concentrating item difficulties at levels along the ability continuum appropriate to each individual's ability. The platykurtosis of conventional test scores may be due to the fact that the test was not perfectly peaked.

Two-stage 2 provided scores that were more reliable than were scores obtained from the conventional test or from Two-stage 1. However, all three reliability estimates were low, ranging from .76 for Two-stage 1 to .83 for Two-stage 2. This is perhaps due to the fact that the method of estimating reliability, the correlation between two parallel forms with no time interval between administrations, includes fewer sources of systematic variance which are included with the score variance instead of with error variance than do most methods of estimating reliability. For example, the reliability coefficients obtained in the present study can be compared with the test-retest stability coefficients of Two-stage 1 and the same conventional test, as studied in Betz & Weiss (1973). The stability of Two-stage 1 was .88 and that of the conventional test was .89. While no stability data is yet available for Two-stage 2, it is reasonable to infer that, given its higher parallel forms reliability (which was actually determined through re-administration of the same test), it would be substantially more stable than either Two-stage 1 or the conventional test. Thus, the correlations between scores obtained from two simulated administrations of the same test are lower than those obtained from the test-retest design with an interval of about five to

six weeks between administrations in the empirical study.

This result can be attributed to 1) the fact that "error" in simulated test responses is entirely random and does not contain some stable item-specific variance and 2) the absence of memory effects. Effects of memory on test-retest stability were found by Betz and Weiss (1973); the stability correlation for individuals who had taken the same measurement test on retesting (thus repeating all 40 items) was .93, as opposed to the value of .88 found for the group as a whole, many of whom had taken a different measurement test on retesting. Similar memory effects were found by Larkin and Weiss (1974a) in a study comparing conventional tests and pyramidal adaptive tests.

Thus, the reliability values obtained in the present study can be considered lower-bound estimates of the stability of test results obtained from two administrations of the same test, where knowledge that is stable but specific to particular item content does not enter into the stability of obtained scores and where the responses of an individual are not affected by previous measurement of the ability (i.e., memory). It should be noted that the obtained parallel forms reliability coefficients are also lower-bound estimates of the internal consistency reliability of the tests (Guilford, 1954; Stanley, 1971).

The relationship between two-stage and conventional test scores was relatively high (.78 to .82) and primarily linear, although Two-stage 2 showed the higher relationship to the conventional test scores. These data indicate that although a majority of variance is shared by the two testing strategies (two-stage and conventional), 33-39% of the variance of either strategy is left unaccounted for.

Ability estimates yielded by Two-stage 2 showed a higher relationship to underlying ability (r=.91) than did ability estimates yielded by Two-stage 1 (r=.87) or the conventional test (r=.90 when based on the sample of 10,000 and r=.89 based on the mean of the sampling distribution of 100 coefficients). It is interesting to note that the correlations between test scores and underlying ability are equal to the squares of the reliability coefficients, which is the prediction yielded by psychometric theory (Gulliksen, 1950). Thus, the square of .91, the correlation between Two-stage 2 scores and ability, is .83, the reliability of Two-stage 2. The reliability of the conventional test (.80) is between $.89^2$(=.79) and $.90^2$(=.81), and the reliability of Two-stage 1 (.76) is equal to the square of its correlation with ability ($.87^2$=.76).

The findings regarding the information or relative precision of measurement at various points along the ability continuum support the conclusion that two-stage testing strategies can provide greater comparability of the precision of ability estimates for individuals at all levels of ability represented in a given population. Two-stage 1 yielded approximately

horizontal information functions, with ability estimates for individuals whose ability levels fell three or more standard deviations from the mean being nearly as precise as those for individuals of average ability. Further, the average level of information provided by Two-stage 1 was greater than that provided by the conventional test and only somewhat less than that provided by Two-stage 2, which yielded the highest level of overall information. Two-stage 2, while showing a loss in precision at the extremes, yielded more constant levels of information than did the conventional test (as indicated by the smaller standard deviation of information values).

The failure of Two-stage 2 to yield a horizontal information function may be due to the strategy used in constructing the test. The average difficulties of the Two-stage 2 measurement tests were chosen to be closer to median ability than were those in the Two-stage 1 measurement tests (see Table 1); this was done in an attempt to maximize the appropriateness of each measurement test for the group of individuals assigned to it. The attempt was found to be successful in an empirical study of Two-stage 2 (Larkin & Weiss, 1974b). Thus, Two-stage 2 was composed of measurement tests more appropriate for individuals near the group mean and less appropriate for individuals whose abilities were near the extremes. The result was a test which did not have the approximately horizontal information function found by Lord (1971c) in his theoretical studies, or by the Two-stage 1 test in this study. Rather, the Two-stage 2 information function was more similar in shape to that of a conventional test, but at a higher level. It would appear that for two-stage tests, just as for peaked conventional tests, the advantages of maximizing the appropriateness of item difficulty for a group of individuals are offset somewhat by a loss in the precision of measurement for individuals whose abilities are not near the mean of the target group.

The finding that Two-stage 2 provided more precise measurement than the conventional test must be interpreted with caution because the average discriminating power of the items used in constructing Two-stage 2 (mean a=.63, mean r_b=.52) was slightly higher than that for the conventional test items (mean a=.54, mean r_b=.47). The results of the present study contradict Lord's findings from a variety of theoretical studies (Lord, 1970, 1971 a,b,c) showing that a conventional test will always provide more information for testees at the mean of the ability distribution than will any adaptive test. But Lord's findings were based on the use of hypothetical, ideal items which were all of the same discriminating power; thus, the relative discriminating power of the items did not influence the superiority of any particular testing strategy. It will be necessary to examine the information-providing characteristics of a conventional test with items as discriminating as those used in Two-stage 2 before it can be concluded that the two-stage test can provide more accurate measurement around the mean of the ability

distribution. However, the superiority of Two-stage 1 to the conventional test with increasing divergence from the mean ability level and its higher overall level of precision of measurement cannot be attributed to differential item discriminating power (the items in Two-stage 1 had a mean a=.55 and a mean r_b=.47, almost identical to that of the conventional test) but is instead attributable to the process of adapting item difficulties to the characteristics of each individual testee.

The results of the simulation studies described here reflected quite closely the results of the parallel empirical study (Betz & Weiss, 1973) with regard to characteristics of the score distributions and the degree and nature of the relationship between two-stage and conventional test scores. The correspondence of the reliability coefficients to the squared correlations between test scores and underlying ability and the similarity of the conventional test and Two-stage 1 information functions to those found in Lord's theoretical studies using restrictive assumptions of "ideal" items are further evidence as to the validity and utility of the simulation model used in this study. Thus, it is concluded that further simulation studies both parallelling and extending the on-going empirical research will be useful in exploring the measurement characteristics of variations of the two-stage testing strategy. For example, most studies of two-stage testing to date have used an even number (usually 4) of measurement tests; in the present study there were four, two at difficulty levels above the mean and two at difficulty levels below the mean. Thus, individuals at the mean ability level for whom the routing test (or any conventional test peaked at the mean ability level) is most appropriate are routed up or down into a somewhat less appropriate measurement test. Using an odd number of measurement tests, where one is peaked at the mean ability level and the others are distributed above or below it, would very likely yield two-stage test scores that were more precise at the mean ability level than conventional test scores given equally discriminating items.

Another approach to improving two-stage testing procedures would involve using more measurement tests with fewer items. However, the narrower the range of routing test scores used to assign individuals to measurement tests, the greater the likelihood that small errors in the estimation of an individual's ability from routing test scores will lead to mis-routing, or routing to an inappropriate measurement test. The possibility of routing errors is probably the major disadvantage of two-stage testing strategies as they are currently being studied, and significant improvements in the procedure would probably result if individuals who had been mis-routed were identified early in the administration of the measurement test and re-routed to a more appropriate test. A recovery routine of this type could

easily be accommodated into the computer administration of two-stage tests and seems to be a necessary and fruitful direction for further investigations of the strategy.

The results of the present study also help clarify criteria which can be used to compare adaptive and conventional strategies. Since the reliability coefficients were shown to be a transformation of the correlation of test scores and ability, they are appropriate criteria for comparison of strategies. However, both reliability coefficients and ability-test score correlations showed only small differences between the strategies. Information functions, on the other hand, showed considerable gains in precision for the adaptive strategy in regions of the ability distribution. When it is not possible to compute information functions, such as in a live-testing study, the present results suggest that differences in reliability coefficients might parallel similar differences in average level of the information functions.

Summary

The improvements made in the construction of Two-stage 2 were reflected by results showing that, in comparison to both Two-stage 1 and the conventional test, scores yielded by Two-stage 2 better reflected the underlying normal distribution of ability, were more reliable, and had a higher relationship to underlying ability. However, although the overall level of information provided by Two-stage 2 exceeded that of the conventional test at all ability levels and that of Two-stage 1 at ability levels within two standard deviations of the mean, it failed to yield the horizontal information function that was predicted and was found for Two-stage 1. Further research is needed to determine the conditions under which two-stage tests will yield horizontal information functions whose values equal or exceed those of conventional tests even at average levels of ability.

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.

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

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.

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 (first edition, 1957).

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

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

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

Larkin, K. C. & Weiss, D. J. An empirical comparison of two-stage and pyramidal adaptive ability testing. Research Report 74-X, Psychometric Methods Program, Department of Psychology, University of Minnesota (in preparation). (b)

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

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

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

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

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

Weiss, D. J. Strategies of adaptive ability measurement. Research Report 74-X, Psychometric Methods Program, Department of Psychology, University of Minnesota, 1974 (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, 1973.

Appendix A

Normal ogive difficulty (b) and discrimination (a) parameters for items of Two-stage 1, Two-stage 2, and the conventional test

Table A-1

Two-stage 1 Item Parameters

Routing Test			Measurement Test 1			Measurement Test 2			Measurement Test 3			Measurement Test 4		
Item Reference Number	b	a	Item Reference Number	b	a	Item Reference Number	b	a	Item Reference Number	b	a	Item Reference Number	b	a
380	-.24	1.00	290	3.38	.42	337	.73	.98	332	-.81	.76	227	-1.63	.71
161	-.26	.86	662	1.93	.57	215	.65	.48	287	-1.27	.44	311	-1.83	.66
104	-.40	.68	328	2.31	.54	291	1.31	.44	33	-.85	.64	189	-1.60	.66
143	-.57	.77	312	3.22	.42	231	.79	.45	110	-1.04	.58	106	-2.39	.62
154	-.58	.66	303	3.70	.38	367	.98	.38	53	-1.01	.52	232	-1.70	.59
365	-.56	.66	627	2.67	.42	397	.83	.37	222	-1.02	.54	76	-2.19	.56
313	-.80	.69	237	3.77	.36	341	.75	.37	155	-1.35	.34	641	-1.89	.52
46	-.81	.66	665	1.01	.90	286	1.16	.32	157	-1.08	.32	93	-2.68	.48
203	-.84	.65	174	1.16	.64	238	.65	.43	239	-1.10	.77	643	-2.56	.44
292	-.58	.48	254	1.16	.59	139	.19	.42	149	-.91	.67	649	-2.21	.44
			288	1.11	.56	324	.09	.37	640	-1.47	.67	256	-2.51	.41
Mean	-.56	.71	162	1.17	.52	148	.42	.35	285	-1.42	.72	73	-3.26	.40
S.D.	.22	.14	140	1.30	.52	347	.14	1.07	322	-1.55	.58	151	-3.19	.41
			263	1.38	.51	283	.15	.97	185	-1.18	.57	81	-2.95	.41
			381	1.79	.51	315	.17	.83	646	-1.18	.57	135	-3.34	.40
			378	1.44	.49	301	.08	.76	235	-1.27	.56	255	-2.58	.59
			115	1.88	.45	342	.17	.77	671	-1.31	.52	63	-2.51	.64
			180	2.07	.43	265	.17	.77	648	-1.27	.51	5	-2.50	.68
			274	2.13	.42	60	.24	.66	112	-1.30	.52	187	-3.53	.45
			659	2.26	.35	635	-.36	.43	117	-1.19	.52	261	-3.99	.39
			193	2.37	.34	377	-.23	.43	657	-1.20	.51	77	-4.08	.42
			360	2.18	.34	644	-.04	.43	293	-1.08	.56	69	-3.64	.50
			260	1.47	.34	133	-.09	.41	241	-1.59	.49	100	-3.55	.56
			168	1.36	.37	355	-.58	.40	94	-1.57	.49	17	-3.19	.68
			306	.97	.49	629	-.26	.40	108	-1.71	.47	96	-1.88	1.14
			159	1.24	.36	318	-.36	.40	653	-1.62	.45	27	-1.92	1.23
			114	.65	.77	116	-.38	.38	650	-2.17	.40	134	-2.21	.96
			107	1.21	.35	634	-.37	.36	642	-1.80	.42	158	-2.26	.98
			304	1.00	.42	239	-.04	.35	141	-1.83	.42	126	-2.54	.88
			660	1.01	.41	252	-.34	.32	276	-2.12	.41	206	-2.43	1.01
			Mean	1.81	.47	Mean	.22	.52	Mean	-1.34	.53	Mean	-2.62	.63
			S.D.	.85	.13	S.D.	.50	.22	S.D.	.35	.11	S.D.	.70	.23

Table A-2

Two-stage 2 Item Parameters

Routing Test			Measurement Test 1			Measurement Test 2			Measurement Test 3			Measurement Test 4		
Item Reference Number	b	a	Item Reference Number	b	a	Item Reference Number	b	a	Item Reference Number	b	a	Item Reference Number	b	a
161	-.26	.86	328	2.31	.54	306	.97	.49	145	-.41	.59	204	-1.15	.73
661	-.30	.58	166	2.03	.64	321	.79	.63	292	-.58	.48	94	-1.57	.49
670	-.28	.62	309	2.47	.48	660	1.01	.40	382	-.48	.64	642	-1.80	.42
52	-.28	.61	298	2.62	.43	114	.65	.77	205	-.62	.47	109	-1.06	.89
599	-.23	.81	627	2.67	.42	630	-.05	1.31	207	-.53	.60	515	-1.33	.62
329	-.21	.86	662	1.93	.57	231	.79	.45	137	-.74	.40	141	-1.83	.42
144	-.18	.63	385	2.35	.42	656	.71	.44	46	-.81	.67	108	-1.71	.47
50	-.23	.50	336	2.04	.49	215	.65	.48	203	-.84	.65	87	-1.10	.99
369	-.22	.56	297	2.31	.40	651	.49	.56	33	-.85	.64	276	-2.12	.41
272	-.13	.98	274	2.13	.42	296	.34	.91	53	-1.01	.52	43	-1.21	.90
			180	2.07	.43	666	.42	.55	188	-.47	.71	214	-2.08	.42
Mean	-.23	.70	245	2.32	.38	375	.46	.49	365	-.56	.66	640	-1.47	.67
S. D.	.05	.16	381	1.79	.50	111	.46	.48	234	-.69	.51	285	-1.42	.71
			273	1.79	.49	340	.30	.78	154	-.58	.66	36	-1.08	1.23
			319	1.49	.62	302	.37	.50	208	-.68	.58	637	-1.40	.75
			359	1.54	.58	271	.33	.53	156	-.63	.65	47	-1.31	.87
			115	1.88	.45	264	.21	.86	270	-.52	.86	232	-1.70	.59
			360	2.18	.34	60	.24	.66	143	-.57	.77	173	-1.43	.76
			652	1.33	.60	113	.25	.61	667	-.73	.57	641	-1.89	.52
			152	1.40	.55	283	.15	.97	211	-.72	.61	189	-1.60	.66
			378	1.44	.49	265	.17	.77	224	-.79	.54	649	-2.21	.44
			263	1.38	.51	386	.14	.70	91	-.59	.83	103	-1.34	.89
			120	1.07	.72	146	.00	.61	37	-.69	.67	88	-1.75	.63
			174	1.16	.64	633	-.08	.50	390	-.73	.63	227	-1.63	.71
			140	1.30	.52	568	-.08	.91	221	-.74	.65	86	-1.55	.77
			288	1.11	.56	59	.17	.64	307	-.84	.56	40	-1.34	1.02
			162	1.17	.52	315	.17	.83	58	-.96	.48	199	-1.42	.92
			337	.73	.98	342	.17	.77	588	-.89	.53	95	-2.20	.50
			294	.79	.70	266	.16	.86	155	-1.35	.34	311	-1.83	.66
			299	.98	.52	347	.14	1.07	535	-.68	.86	643	-2.56	.44
			Mean	1.72	.53	Mean	.35	.68	Mean	-.71	.61	Mean	-1.60	.68
			S.D.	.56	.13	S.D.	.30	.21	S.D.	.19	.12	S.D.	.37	.21

Table A-3

Item Parameters for the Conventional Test

Item Reference Number	Difficulty (b)	Discrimination (a)
58	-.96	.48
221	-.74	.65
307	-.84	.56
393	-.95	.49
211	-.72	.61
224	-.78	.54
390	-.73	.63
667	-.73	.57
156	-.63	.65
208	-.68	.58
234	-.69	.51
52	-.28	.61
137	-.74	.40
176	-.90	.34
207	-.53	.60
218	-.93	.33
205	-.62	.47
382	-.48	.64
391	-.53	.48
626	-.29	.65
645	-.32	.50
661	-.30	.58
670	-.28	.62
327	-.25	.57
50	-.23	.50
144	-.18	.63
369	-.22	.56
233	-.17	.47
636	-.15	.54
633	-.08	.50
146	.00	.61
295	-.04	.47
113	.25	.61
267	.19	.44
59	.17	.64
271	.33	.53
302	.37	.50
375	.46	.49
666	.42	.55
651	.49	.56
Mean	-.33	.54
S. D.	.43	.08

Appendix B

Routing test scores and corresponding initial ability estimates used in the assignment of testees to Two-stage 2 measurement tests.

Routing Test Score (number correct)	Ability Estimate (standard scores)	Mean Difficulty Level of Assigned Measurement Test
2.5*	-2.45	-1.6 (Test 4)
3	-1.90	-1.6 (Test 4)
4	-1.20	-1.6 (Test 4)
5	-.69	-.71 (Test 3)
6	-.23	-.71 (Test 3)
7	.23	.35 (Test 2)
8	.75	.35 (Test 2)
9	1.44	1.73 (Test 1)
9.5*	1.99	1.73 (Test 1)

*Ability estimates are infinite for perfect scores (10 correct) or for scores at or below chance level ($\leq$ 2 correct).

Appendix C

Description of the algorithm for SIMTEST, the computer program controlling simulated test administration.

Program SIMTEST is written to generate hypothetical ability and four test scores for each of 100 "testees" on each run of the program. Program SIMTEST, written in FORTRAN for a Control Data Corporation 6400 computer, runs in a time-shared mode and proceeds as follows:

1. Normal ogive difficulty (b) and discrimination (a) parameters are read for each item in the item pool.
2. An initialization value ("seed") is read for the random number generator.
3. Using the "seed," 100 ability levels are generated from a theoretical normal distribution using Subroutine NORMAL (a University of Minnesota Computer Center systems subroutine). This subroutine is a pseudo-random generator of real numbers from a normal distribution with mean 0 and variance 1.
4. Subroutine RAN2F (University of Minnesota Computer Center) is used to generate 160 random numbers from a rectangular distribution of real numbers between 0 and 1. These 160 numbers are stored for use in subroutine ITEMSYM (Step 7).
5. The ability level of the hypothetical subject is sent to one of the testing subroutines (Two-stage 1, Two-stage 2, or conventional), where the determination of the first or next item to be administered is made.
6. In Subroutine ITEMSYM, the parameters of that item--a_i, b_i, and c_i (the guessing parameter set at .2)--and the individual's ability level Θ are entered into the following equation:
$$P_i(\Theta) = c_i + (1 - c_i)\,\Phi[a_i\,(\Theta - b_i)]$$
The result, $P_i(\Theta)$, is the probability that a person with ability Θ will answer item i correctly.
7. In Subroutine ITEMSYM, $P_i(\Theta)$ is compared to the random number generated in Step 4 which corresponds to the order of administration of item i. Thus, the value of the random number p_i is compared to the probability $P_i(\Theta)$ that the individual will answer item i correctly.
   If $P_i(\Theta) > p_i$, the item is scored "correct."
   If $P_i(\Theta) < p_i$, the item is scored "incorrect."
   The values of p_i and $P_i(\Theta)$ occur with sufficient places to the right of the decimal point that the chance of $p_i = P_i(\Theta)$ is extremely small and, in fact, has not occurred.
8. The dichotomized item response is returned to the testing program, which stores it and may use it to determine the next item to be administered. After the administration of each 40-item test, the total score for that test is calculated for that "testee."

In order to generate a rectangular distribution of underlying ability, the procedure described in Step 3 was replaced by a procedure in which a particular level of ability was read in and used as the underlying ability level of all 100 "testees" simulated in that run.

Appendix D

Unaveraged values of the information function for two-stage and conventional tests, from Time 1 and Time 2 administrations.

Table D-1

Information values for Time 1 and Time 2 administrations of two-stage and conventional tests ("equal-frequency" distribution, N = 100 at each level of θ)

Ability (θ)	Information ($I_x(\theta)$)					
	Two-stage 1		Two-stage 2		Conventional	
	Time 1	Time 2	Time 1	Time 2	Time 1	Time 2
3.2	1.88	3.14	1.37	2.68	1.06	.65
3.0	2.44	2.41	1.61	1.85	.03	.01
2.5	2.82	4.24	3.47	3.46	1.07	1.18
2.0	4.00	4.18	3.34	4.62	3.37	3.22
1.5	3.33	3.85	2.92	5.84	3.86	4.90
1.0	3.16	2.74	4.83	5.08	4.81	4.03
.5	3.46	2.78	5.07	6.36	6.07	4.53
.1	2.87	2.46	6.93	5.51	3.96	4.96
-.1	3.09	2.62	5.36	4.14	4.31	4.45
-.5	3.95	3.25	4.07	5.75	4.88	3.62
-1.0	3.85	3.87	5.35	5.09	3.71	3.34
-1.5	2.63	4.70	3.36	5.23	3.51	2.51
-2.0	2.64	3.73	2.81	2.35	2.33	2.66
-2.5	2.60	2.26	2.21	3.66	1.38	1.25
-3.0	1.75	3.07	.75	1.51	.28	.51
-3.2	1.64	2.55	.58	1.00	.08	.22
Mean	2.88	3.24	3.38	4.01	2.79	2.63
S. D.	.74	.76	1.82	1.70	1.92	1.76

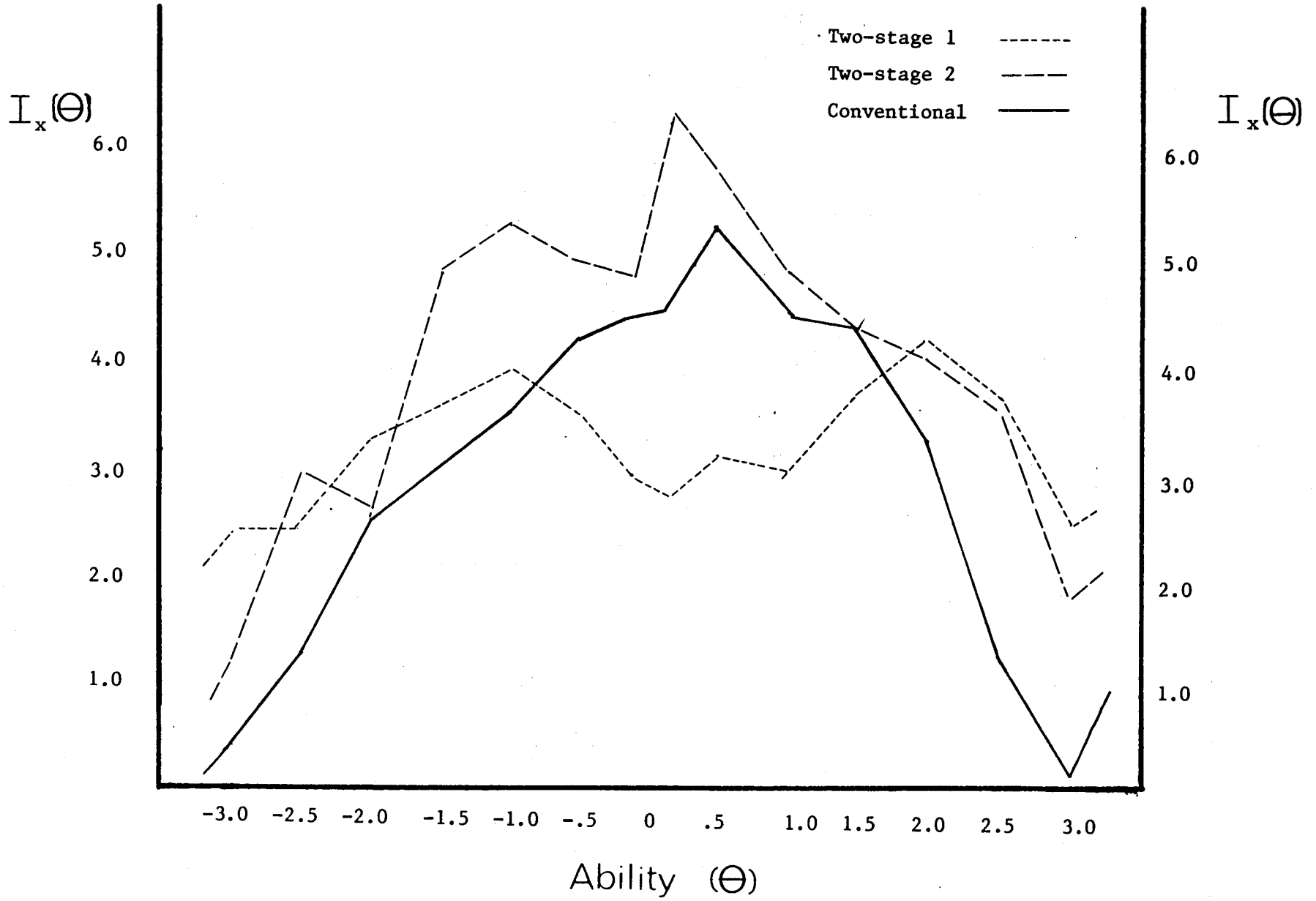
Table D-2

Information values for Time 1 and Time 2 administrations of two-stage and conventional tests (normal distribution of underlying ability, total N = 10,000)

Interval of Ability (Θ)	Information ($I_x(\Theta)$)					
	Two-stage 1		Two-stage 2		Conventional	
	Time 1	Time 2	Time 1	Time 2	Time 1	Time 2
3.1 to 3.3	2.94	6.08	1.12	.51	.58	**
2.9 to 3.1	3.06	13.65*	1.46	1.58	.04	.42
2.7 to 2.9	3.64	4.92	2.22	3.16	.12	.01
2.5 to 2.7	2.23	2.35	4.73	3.10	.43	.41
2.3 to 2.5	3.08	3.98	4.48	2.74	1.93	1.40
2.1 to 2.3	3.57	4.23	2.98	4.46	2.02	1.85
1.9 to 2.1	2.64	3.64	3.46	4.03	4.23	2.73
1.7 to 1.9	4.00	3.95	3.04	4.38	3.48	3.99
1.5 to 1.7	3.49	3.95	3.96	3.94	3.46	3.64
1.3 to 1.5	3.40	3.42	4.21	5.19	3.97	4.54
1.1 to 1.3	3.60	3.93	5.12	5.26	4.05	4.35
.9 to 1.1	3.49	3.52	4.65	5.00	4.36	5.19
.7 to .9	3.04	3.59	4.66	4.95	4.27	4.45
.5 to .7	3.23	2.94	4.13	4.27	4.54	4.29
.3 to .5	2.88	2.94	4.60	4.90	4.78	4.96
.1 to .3	2.55	3.04	5.07	5.02	4.58	4.53
-.1 to .1	3.14	3.01	4.87	5.74	5.18	4.39
-.3 to -.1	2.80	3.21	5.27	5.94	4.37	4.51
-.5 to -.3	3.18	2.98	5.52	5.50	4.26	4.24
-.7 to -.5	3.43	3.71	5.48	5.82	4.28	3.63
-.9 to -.7	3.07	3.33	5.09	5.63	3.87	3.70
-1.1 to -.9	3.12	3.32	5.09	5.44	3.64	3.33
-1.3 to -1.1	3.10	3.71	4.51	5.03	3.78	3.33
-1.5 to -1.3	2.97	4.02	3.79	4.70	3.85	2.87
-1.7 to -1.5	3.32	3.82	4.17	4.61	3.39	2.80
-1.9 to -1.7	2.99	3.85	4.30	3.71	2.98	2.50
-2.1 to -1.9	3.03	3.53	3.09	3.54	2.97	2.11
-2.3 to -2.1	3.71	4.09	2.41	3.17	2.57	1.40
-2.5 to -2.3	2.39	3.56	1.93	1.87	1.46	1.50
-2.7 to -2.5	3.21	3.74	3.59	1.86	3.10	.94
-2.9 to -2.7	3.13	4.68	1.44	1.26	1.26	.33
-3.1 to -2.9	2.37	5.11	.32	2.07	.71	.98
-3.3 to -3.1	4.09	3.18	.62	7.00	.29	.01
Mean	3.15	4.03	3.68	4.10	2.99	2.79
S. D.	.43	1.87	1.48	1.55	1.56	1.63

*Score variance was extremely small; deleting this value results in a mean of 3.73 and a variance of .72.

**Value was infinite because there was no variance (the two scores falling in this interval were equal).



Appendix E, Figure E-1. Information functions of two-stage and conventional tests using "equal-frequency" distribution of ability (based on results presented in Table 7).

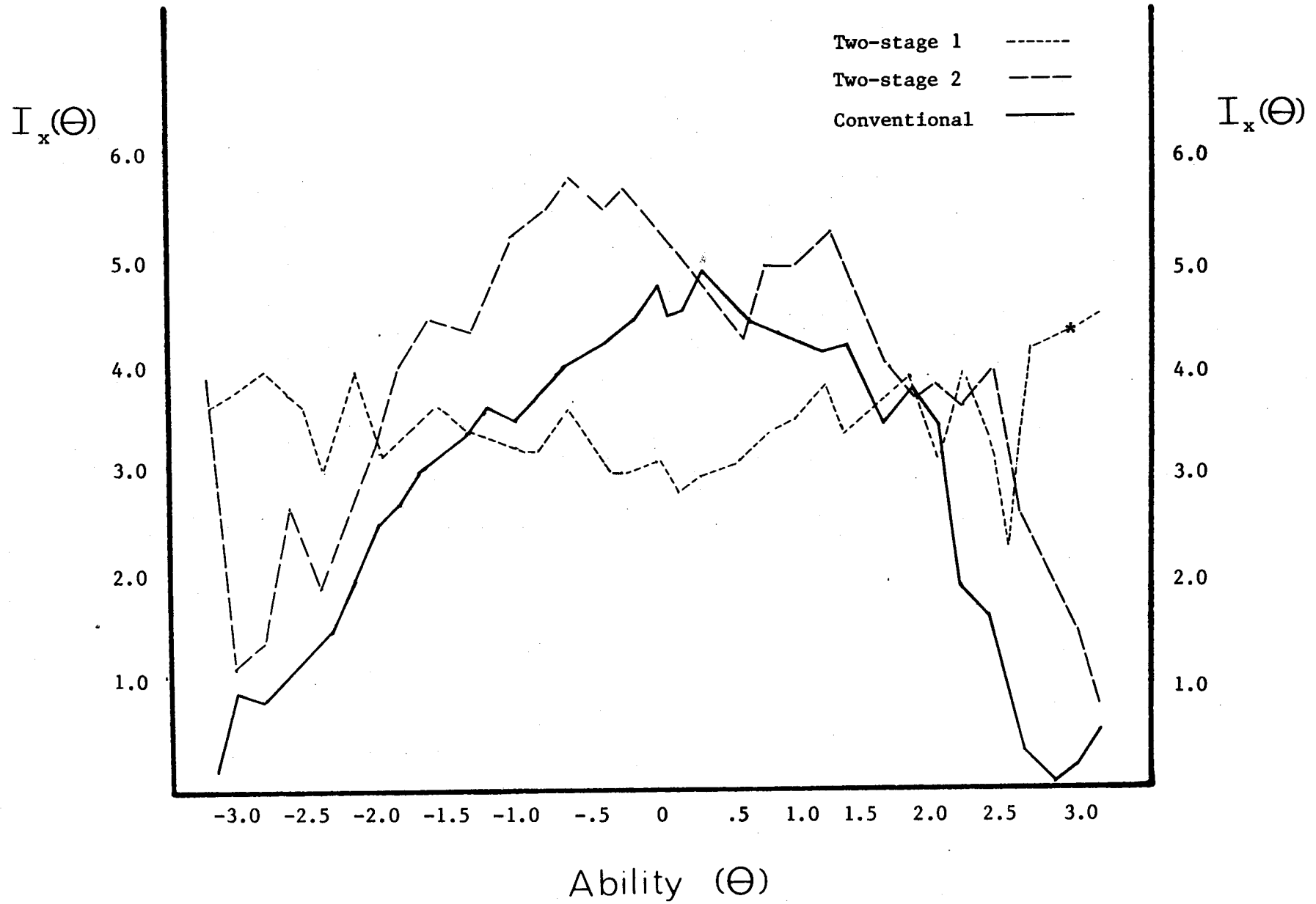


Figure E-2. Information functions of two-stage and conventional tests using normal distribution of ability (based on results presented in Table 8).

*$I_x(0)$ = 8.35; variance was unusually small.