DOCUMENT RESUME

ED 111 861 TM 004 828

AUTHOR Betz, Nancy E.; Weiss, David J.
TITLE Empirical and Simulation Studies of Flexilevel Ability Testing. Research Report No. 75-3.
INSTITUTION Minnesota Univ., Minneapolis. Dept. of Psychology.
SPONS AGENCY Office of Naval Research, Washington, D.C. Personnel and Training Research Programs Office.
REPORT NO RR-75-3
PUB DATE Jul 75
NOTE 56p.
AVAILABLE FROM Psychometric Methods Program, Dept. of Psychology, University of Minnesota, Minneapolis, Minnesota 55455 (while supplies last)

EDRS PRICE MF-$0.76 HC-$3.32 Plus Postage
DESCRIPTORS *Ability; College Students; Comparative Analysis; *Computer Oriented Programs; Feedback; Individual Differences; Item Banks; Measurement Techniques; Memory; *Response Style (Tests); *Simulation; Test Construction; *Testing; Test Reliability
IDENTIFIERS *Flexilevel Test

ABSTRACT

A 40-item flexilevel test and a 40-item conventional test were compared using data obtained through (1) computer-administration of the two tests to three groups of college students, and (2) monte carlo simulation of test response patterns. Results indicated the flexilevel score distribution better reflected the underlying normal distribution of ability, and that the flexilevel test had a higher paralleled-forms reliability and a higher relationship to underlying ability level than did the conventional test. The overall test-retest stability of the two tests was equivalent, but there was evidence indicating that memory effects inflated the stability of the flexilevel test scores less than that of conventional test scores. The flexilevel provided more accurate measurement at almost all ability levels, although its information function was similar in shape to that of the conventional test. However, the interpretation of differences in the level of information provided were confounded by differences in the average discriminating power of the items in the two tests. The flexilevel test also appeared to reduce random guessing behavior in comparison to the conventional test. (Author)

TM

EMPIRICAL AND SIMULATION STUDIES OF FLEXILEVEL ABILITY TESTING

Nancy E. Betz

and

David J. Weiss

Research Report 75-3

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

July 1975

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

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER

Research Report 75-3

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)

Empirical and Simulation Studies of Flexilevel Ability Testing

5. TYPE OF REPORT & PERIOD COVERED

Technical Report

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)

Nancy E. Betz and David J. Weiss

8. CONTRACT OR GRANT NUMBER(s)

N00014-67-0113-0029

9. PERFORMING ORGANIZATION NAME AND ADDRESS

Department of Psychology
University of Minnesota
Minneapolis, Minnesota 55455

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

P.E.:61153N PROJ.:RR042-04
T.A.:RR042-04-01
W.U.:NR150-343

11. CONTROLLING OFFICE NAME AND ADDRESS

Personnel and Training Research Programs
Office of Naval Research
Arlington, Virginia 22217

12. REPORT DATE

July 1975

13. NUMBER OF PAGES

46

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

15. SECURITY CLASS. (of this report)

Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

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

18. SUPPLEMENTARY NOTES

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

testing
ability testing
computerized testing
adaptive testing
sequential testing
branched testing
individualized testing
tailored testing
programmed testing
response-contingent testing
automated testing

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

A 40-item flexilevel test and a 40-item conventional test were compared using data obtained through 1) computer-administration of the two tests to three groups of college students, and 2) monte carlo simulation of test response patterns. Results indicated the flexilevel score distribution better reflected the underlying normal distribution of ability, and that the flexilevel test had a higher parallel-forms reliability and a higher relationship to underlying ability level than did the conventional test.

The overall test-retest stability of the two tests was equivalent, but there was evidence indicating that memory effects inflated the stability of the flexilevel test scores less than that of conventional test scores. The flexilevel test provided more accurate measurement at almost all ability levels, although its information function was similar in shape to that of the conventional test. However, the interpretation of differences in the level of information provided were confounded by differences in the average discriminating power of the items in the two tests. The flexilevel test also appeared to reduce random guessing behavior in comparison to the conventional test.

Contents

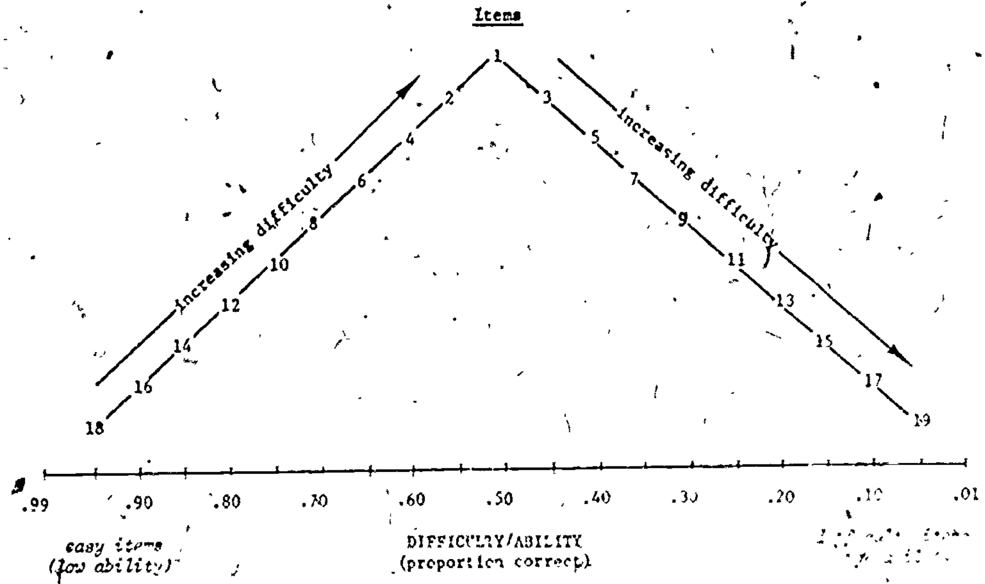
EMPIRICAL AND SIMULATION STUDIES OF FLEXILEVEL ABILITY TESTING

One result of the growing sophistication and availability of time-shared computer facilities has been increased interest in new modes of testing and instruction. In the area of ability measurement, much research has been directed at investigating various strategies of tailored (Lord, 1970) or adaptive (Weiss & Betz, 1973) ability testing. The general aim of adaptive testing procedures is to "adapt" or "tailor" the difficulty level of the items presented to the ability level of an individual as estimated from item response patterns. Consequently, as testing proceeds, the items administered will be increasingly appropriate for the accurate measurement of that individual's ability.

Adaptive testing strategies are differentiated by the set of rules used to select items during the testing procedure (Weiss, 1974). The most extensively researched adaptive strategy is the pyramidal or "tree-structure" model. This approach uses a branching (or item selection) rule in which, following a correct response to an item, the examinee receives a slightly more difficult item, and following an incorrect response, the examinee receives a slightly less difficult item. Research to date, summarized by Weiss & Betz (1973) and Larkin & Weiss (1974), has shown that pyramidal strategies can yield equal or better reliability and validity than conventionally-administered tests while requiring substantially fewer items to be administered. The flexilevel test (Lord, 1971b) is a modification of the pyramidal strategy which would permit paper and pencil administration and which would require a smaller initial item pool than is required by pyramidal strategies.

Figure 1 illustrates the item structure for a flexilevel test. As Figure 1 shows, the flexilevel test consists of one item at each of a number of equally-spaced difficulty levels. Item 1 in Figure 1 is an item of approximately median difficulty (p=.50). The even-numbered items decrease in difficulty with increasing distance from the median difficulty level, while the odd-numbered items increase in difficulty.

Figure 1
ITEM STRUCTURE FOR A TEN-STAGE FLEXILEVEL TEST



In the flexilevel test illustrated, ten items would be administered to each individual. The total item structure requires 19 items or, in general, 2N-1 items, where N is the number of items to be administered to each individual. The first item administered to all individuals is the median difficulty item (item 1) for the group taking the test. Following administration of the first item, a differential branching rule determines item selection: following a correct response to an item, the examinee receives the next more difficult item previously unanswered; following an incorrect response, the examinee receives the next less difficult item previously unanswered.

How the flexilevel test adapts item difficulties to individual differences in ability level can be seen by an examination of the examples shown in Figure 2. Figure 2a illustrates the path through a flexilevel test for an examinee of relatively high ability. All testees begin with item 1, an item of median difficulty. Each correct answer leads to an item of higher difficulty; thus, correct answers to items 1, 3, 5, 7, 9 and 11 led to the administration of progressively more difficult items, moving from an item at p=.50 to one at p=.20. Item 13 was answered incorrectly, and the next *less* difficult item not already administered was item 2, with difficulty p=.55. Item 2 was answered correctly, and the next more difficult item not already administered was item 15, with difficulty p=.15. Following an incorrect response to this item, item 4, with difficulty p=.60, was administered. Thus, this examinee received ten items in the difficulty range of p=.60 to p=.15.

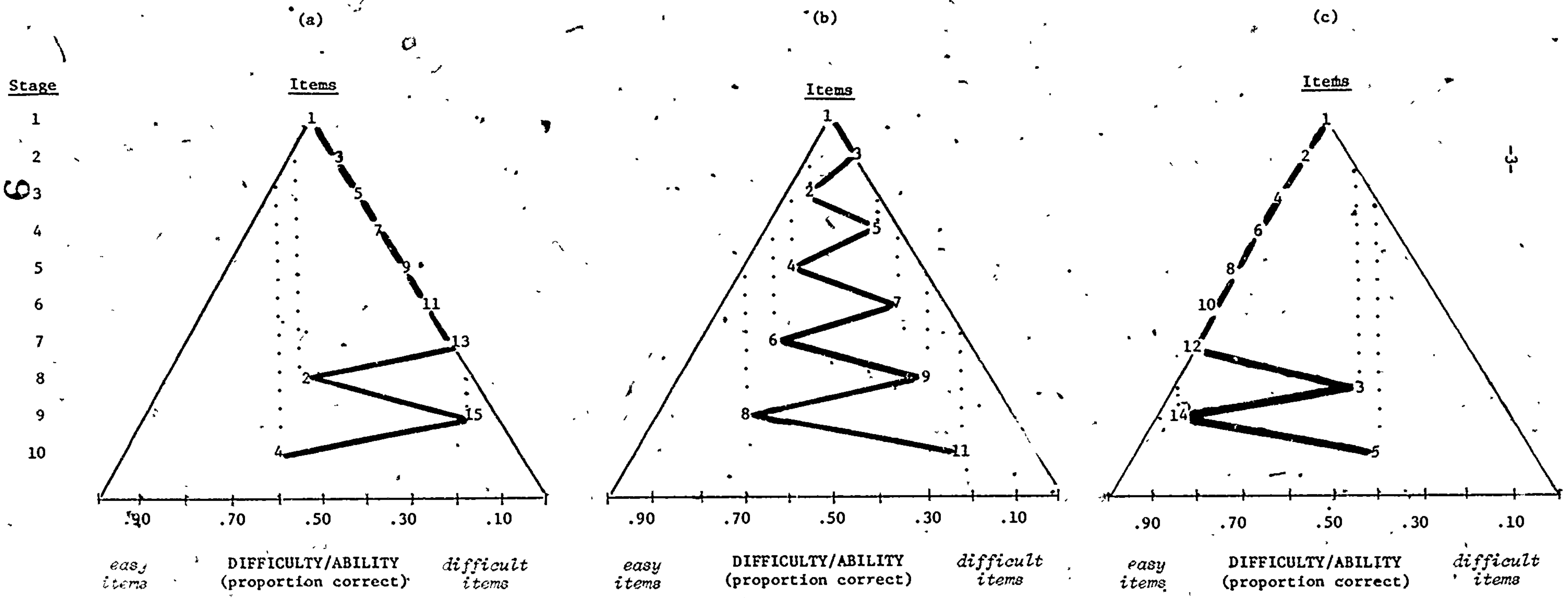
Figure 2b shows how an examinee of average ability might move through the item structure, alternating between successively more difficult and successively less difficult items. Since this examinee is of average ability, the odd-numbered items (except for item 1) are too difficult for him, and he answers them incorrectly; the even-numbered items are too easy for him, and he answers them correctly. Thus, an examinee of average ability might be administered ten items in the difficulty range of p=.70 to p=.25.

Finally, Figure 2c illustrates the path that might be taken by an examinee of relatively low ability. Incorrect answers to items 1, 2, 4, 6, 8 and 10 lead to the administration of progressively less difficult items, culminating in the administration of item 12, with difficulty p=.80. Then, alternating correct and incorrect answers lead to the administration of items at difficulties p=.45, p=.85 and p=.40. Thus, this examinee received ten items in the difficulty range of p=.85 to p=.40.

The flexilevel test can be scored by counting the number of correct responses. Lord (1971b) shows that the greater the number correct, the more difficult was the subset of items answered and, therefore, the higher is the ability level of that examinee. However, Lord also shows that examinees with the same total number correct may be further differentiated according to whether the last item was answered correctly or incorrectly; those who answered the last item incorrectly have answered a more difficult subset of items and have higher ability than those with the *same* total number correct who responded correctly to the last item administered. Accordingly, Lord proposes that an additional half-point be added to the number-correct scores of examinees responding incorrectly to the last item administered.

Figure 2

SAMPLE PATHS THROUGH A TEN-STAGE FLEXILEVEL TEST



In summary, the flexilevel test adapts item difficulties to the ability level of the examinee being tested using a branching procedure which selects from the 2N-1 items available a subset of N items to be administered to that examinee. The N items administered are those whose difficulties are nearest to the examinee's ability level. Because of this adaptive property, the flexilevel test should have several advantages in comparison to conventional ability testing procedures.

First, since examinees will receive fewer items that are much too difficult or much too easy for them, and thus fewer items that are inappropriate for the accurate measurement of their abilities, it is possible that the flexilevel test will yield ability estimates as reliable and valid as those of conventional tests utilizing considerably more items. Stanley (1971) suggests that the *effective* length of a conventionally-administered test is considerably less than the total number of items administered; it is the purpose of an adaptive test to select for administration those items that *are* effective for measuring the ability of a given examinee.

Second, the flexilevel test should provide ability estimates whose reliability and validity are more nearly equivalent for examinees of different ability levels. Several reports (Baker, 1964; Levine & Lord, 1959; Lord, 1957, 1959) have concluded that the precision or reliability of measurement for a given individual is partly dependent on his/her "true score." Thorndike (1951) and Davis (1952), among others, have shown that the standard error of measurement will be minimum for examinees whose ability levels correspond to that point on the ability/item difficulty scale where the item difficulties in the test are concentrated. On the conventional "peaked" ability test, with item difficulties concentrated around p=.50, the error of measurement should be minimum for examinees of average ability and will increase for individuals whose ability levels deviate from the average. Thus, ability estimates for high and low ability examinees will be less reliable than those for average ability examinees. Further differential error in test scores is contributed by differences in the amount of guessing on multiple-choice tests. While guessing reduces the reliability and validity of measurement for all subjects (e.g., Ebel, 1969; Frary & Zimmerman, 1970; Lord, 1957) the increase in error is greatest for low ability subjects. According to Nunnally (1967), on a conventional test where all items are attempted, low ability subjects will guess the most because they know the least. Thus, the flexilevel test, where item difficulties are concentrated around the ability level of each examinee, should yield ability estimates which will tend to be equally reliable across the ability continuum.

Research on Flexilevel Tests

Research to date on flexilevel testing includes one theoretical study (Lord, 1971d), one real-data simulation study (Kocher, 1974) and one live-testing study (Olivier, 1974).

Theoretical study. Lord's (1971d) study comparing the measurement effectiveness of flexilevel and conventional tests was based on the assumptions and mathematics of item characteristic curve theory (Lord & Novick, 1968). The flexilevel tests studied were composed of 60 items (thus requiring a total item structure of 119 items), all having the same discriminating power (normal ogive parameter a equal to .50) and having difficulties distributed along the ability continuum such that the distance between successive item difficulties was a constant, d. The tests were scored using number-correct plus an additional half-point for item response patterns having the last item incorrect. The conventional or "standard" tests used for comparative purposes were composed of 60 equally-discriminating items ($a=.50$). In one of these tests, all items were of median (normal ogive parameter $b=0.0$) difficulty. The other two conventional tests were intended to measure most effectively or be most highly discriminating at two points on the ability continuum. For maximally effective or discriminating measurement at $\theta=\pm2$, one conventional test had 30 items at $b=+2$ and 30 at $b=-2$. For maximally effective measurement at $\theta=\pm3$, the other conventional test had 30 items at $b=+2.8$ and 30 at $b=-2.8$.

Flexilevel and conventional tests were compared in terms of information functions, which indicate the relative accuracy of measurement across the ability continuum. The value of the information function at a given level of ability indicates how well the test scores obtained by individuals of that ability accurately reflect their "true" ability. The greater the value of information at a given level of ability, the more accurate is the measurement or, in other words, the smaller is the confidence interval for estimating true ability from test scores.

Information values are not meaningful in any absolute sense because they are dependent on the scale used to measure ability (θ), but information values calculated from two or more testing strategies assuming the same θ scale can be directly compared, with larger values indicating more accurate measurement. Further, the ratio between the two tests' information values at a given level of ability can be interpreted in terms of the relative numbers of items required to provide equal accuracy of measurement for individuals at that ability level. For example, if, for a given θ level, the information value of one test is twice that of a second test, the first test provides as much information as the second test while requiring half the number of items.

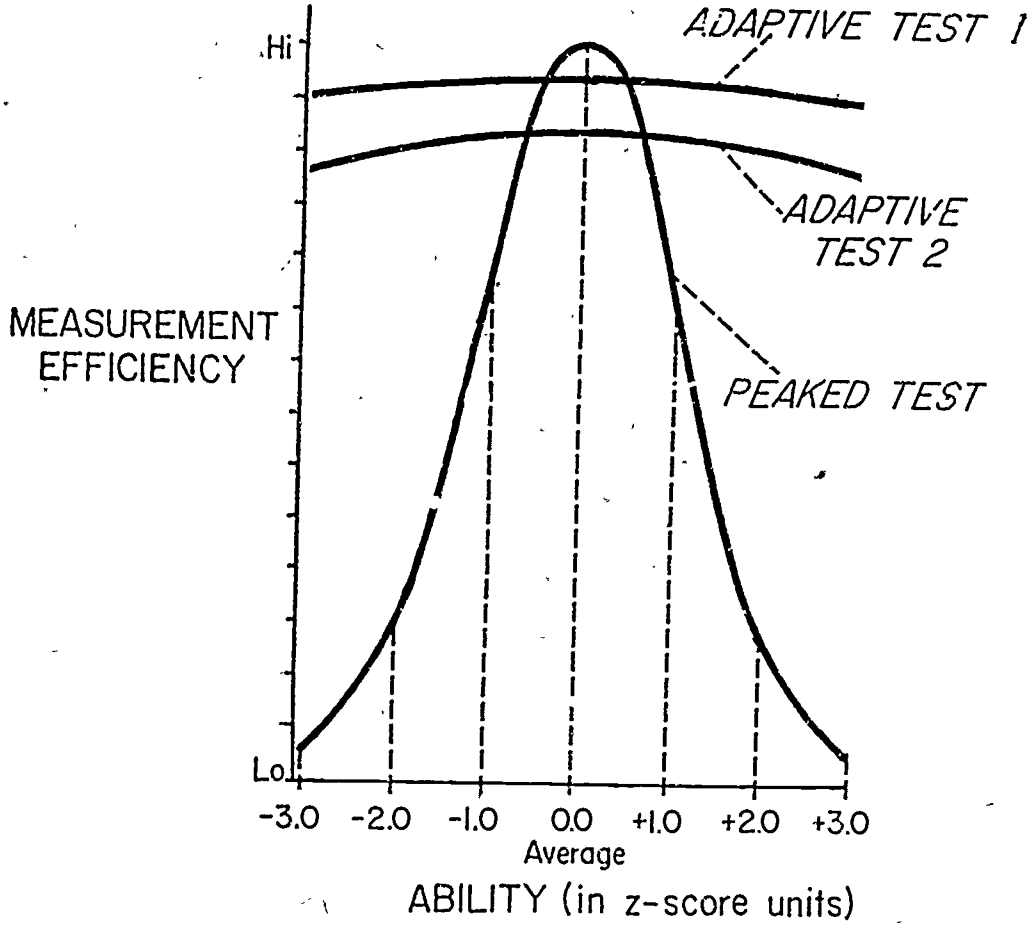
Lord (1971d) found that the flexilevel tests provided more information throughout the ability range than did the conventional tests designed to discriminate at two points ($\theta=\pm2$ or $\theta=\pm3$) on the ability continuum. The conventional test peaked at the median ability level ($b=0.0$) provided more information than did the flexilevel tests at ability levels around the median, but as ability level deviated from the average, the flexilevel tests provided increasingly more information than did the conventional test. For example, the 60-item flexilevel test in which d, the distance between successive item difficulties, was equal to .033/2a (equal to .033 since a was equal to .50) measured as accurately as a 58-item conventional test at $\theta=0$, a 60-item conventional test at $\theta=\pm1$, a 69-item conventional test at $\theta=\pm2$, and an 86-item conventional test at $\theta=\pm3$. Thus, for any examinee with an ability level outside the range of $\theta=\pm1$, the flexilevel test provided more accurate measurement. These results were obtained under the assumption of no guessing.

The results obtained when the guessing parameter "c" was set at .2 were similar to those obtained when no guessing was assumed, except that the superiority of the flexilevel test outside of the range $\theta=\pm1$ was more pronounced for the low ability levels. For example, the flexilevel test measured as accurately as an 83-item conventional test at θ=+3, but as accurately as a 114-item conventional test at θ=-3. Thus, these data indicate that the advantage of flexilevel tests at low ability levels is significantly greater when correct responses are likely as the result of guessing.

Lord's finding that the peaked conventional test provided more information for individuals of near-average ability, while the flexilevel tests provided more information for individuals whose ability levels deviated appreciably from the mean, is in agreement with his other theoretical studies comparing adaptive and conventional tests (Lord, 1970, 1971a,e) In general, the comparative efficiency or precision of measurement of adaptive versus conventional testing strategies as studied theoretically is summarized graphically in Figure 3. Figure 3 illustrates that while the conventional peaked test does provide superior measurement around the mean ability level, the accuracy of measurement of the adaptive tests is more

Figure 3

A HYPOTHETICAL ILLUSTRATION OF THE COMPARATIVE MEASUREMENT EFFICIENCY (PRECISION OR INFORMATION) OF CONVENTIONAL PEAKED AND ADAPTIVE TESTS



constant across all levels of ability and exceeds that of peaked tests beyond a point above and below the mean ability level. The importance of these findings is that they indicate that an individual will be more accurately measured as the items administered to him/her are more appropriate (i.e., nearer in difficulty level) to his/her level of ability.

However, Lord's results concerning the comparative accuracy of measurement of flexilevel (and other adaptive) tests and conventional tests are limited by the assumption of items with equal discriminating power, and having difficulty levels equal to theoretical specifications. It is uncertain whether such results can be generalized to situations in which tests must be constructed using item pools containing finite numbers of items having parameters that can only be estimated and which do not necessarily correspond to ideal specifications. For example, in a simulation study of two-stage adaptive testing procedures using real item parameters (Betz & Weiss, 1974), it was found that one two-stage test provided more information than did a conventional test at all ability levels, including the mean. Although in that study the average discriminating power of the items in the two-stage test was slightly greater than that of the conventional test items, they do suggest some skepticism regarding the generalizability of results obtained under the assumption of theoretically ideal items.

Real-data simulation study. In the study by Kocher (1974), responses to conventional test items were scored as if the tests had been administered using the structure and branching rules of the flexilevel strategy. The study used data from five previously administered conventional tests. Three of these tests, consisting of 42, 36, and 36 items respectively, were classroom examinations administered to 180 college students enrolled in a junior-level course in introductory educational measurements. Pearson product-moment correlation coefficients were calculated between scores on the 21, 18 and 18-item flexilevel tests and scored on the appropriate parent tests. In addition, the correlation between the sum of the standard scores on all three flexilevel tests and the sum of the standard scores on all three parent tests was computed.

The last two conventional tests were semester final examinations in a high school geometry course. The first group, consisting of 412 students, had been administered a 100-item examination. The second group, consisting of 485 students, had been administered a 70-item examination. Again, correlations between the simulated flexilevel scores and scores on the appropriate parent test were calculated.

Results indicated that the correlations between simulated flexilevel scores and scores on the parent tests ranged from .90 to .96; the correlation between the two sets of summated scores obtained in the college group was .96. The size of these correlations, which Kocher interpreted as parallel-forms reliability coefficients, was taken to indicate that flexilevel scores could be validly substituted for conventional test scores and have the advantage of using fewer items.

However, interpreting these correlations as parallel-forms reliability coefficients is not valid because the flexilevel items were a subset of the items in each parent test. The item overlap between the two tests would suggest that the obtained correlation coefficients are artifactually high. In addition, the results of the study are limited by the fact that the flexilevel tests were not actually administered to the examinees. Thus, there was no allowance for the possible psychological effects on an examinee of taking a test in which item difficulty is at least somewhat adapted to his/her ability level.

Live-testing study. A study which employed paper and pencil administration of a flexilevel test was reported by Olivier (1974). In this study, eighth-grade students were first administered the Florida Eighth Grade Test Battery. Approximately one month later, they were administered either a 40-item conventional test or a 20-item flexilevel test. The 39 items needed for the total flexilevel structure were the same items as were used in the 40-item conventional test; these items were taken from the reading vocabulary subtest administered initially as part of the Eighth Grade Test Battery. In order to compare the flexilevel test to a conventional test with the same number of items, three 20-item conventional subtests were extracted from the total 40-item conventional test. The three 20-item tests were constructed by 1) randomly selecting 20 of the 40 items; 2) selecting the even-numbered items; and 3) selecting the 20 items with difficulty values closest to p=.67 (considered the optimal level of difficulty for the group when items were four-alternative multiple choice).

Results showed, first, that the flexilevel test was less internally consistent and, therefore, had a larger standard error of measurement than any of the conventional tests. Second, the flexilevel test showed a lower correlation with an external criterion than did the conventional tests. Third, and the only result favorable to the flexilevel strategy, it was found that item difficulties calculated from the flexilevel administration were closer to p=.67 and had a smaller standard deviation than the item difficulties as calculated on the normative sample. This result indicates that item difficulties were more appropriate for the individuals to whom the items were administered.

However, this study contained several methodological errors which severely limit the fairness of the comparison between flexilevel and conventional testing procedures. First, a one-factor random effects analysis of variance model (Stanley, 1971, pp. 425-428) was used to estimate the internal consistency reliability of the flexilevel test. However, nearly all of the assumptions of this model were violated in the study--an infinitely large item pool from which items are randomly selected for administration to each subject, random assignment of subjects to treatments, and a probability approaching zero that two examinees will attempt the same item. Olivier justifies the violations on the basis of a lack of an alternative method for computing internal consistency reliabilities.

Olivier claimed that the adequacy of the method of reliability estimation was indirectly supported by the fact that the correlation between the flexilevel test and the criterion was lower than that between the conventional tests and the criterion; presumably the lower correlation was due to the attenuation caused by the lower reliability rather than to a lesser proportion of shared variance. However, the criterion itself was questionable on two bases. It consisted of the combined score from six other subtests in the test battery: 1) reading comprehension; 2) reading essential skills; 3) study skills; 4) occupational information; 5) mathematics problem solving; and 6) mathematics essential skills. Only the two reading tests would appear to have any relevance as criteria for the adequacy of a vocabulary test. And regardless of the content of the criterion test, it is questionable whether another conventionally-administered test should be the only standard for evaluating the relative efficiency of conventional versus adaptive testing procedures, since the higher correlation between the two conventional tests could be due to method variance.

Finally, paper and pencil administration of the flexilevel test was found to present several serious difficulties which reduced the accuracy of the test data collected. First, over 10% of the flexilevel protocols had to be discarded because the examinees made errors in following the branching instructions. Second, another 10% of the answer sheets were found to have faulty ink, thus causing many examinees to misroute themselves even though they were following the directions properly. These latter protocols were retained in the analysis with unknown effects on the results. Third, in order to follow the branching rules, examinees knew whether they had answered each item correctly or incorrectly; it is possible that such immediate feedback may have aroused anxiety in examinees given the flexilevel test that was not aroused in examinees administered the conventional test.

The lower reliability of the flexilevel test found in Olivier's study may be related to one potentially disadvantageous characteristic of the test; while the flexilevel test does identify a *region* of the item pool of approximately appropriate difficulty for each examinee, after the *maximally* appropriate difficulty level is reached the remaining items administered tend to be increasingly *divergent* from the examinee's ability level. Reference to Figure 2 provides an illustration of this characteristic. For the high ability examinee in Figure 2a, the most appropriate level of item difficulty probably lies between p=.25 and p=.20. The first seven items administered converge on this level of difficulty, but the items administered following an incorrect response to item 13 are increasingly divergent, having difficulties of p=.55, p=.15 and p=.60. For the average ability examinee in Figure 2b, for whom the median difficulty level is most appropriate, the items administered become progressively less appropriate to his/her ability level.

The net effect of this divergence characteristic is that as the flexilevel test proceeds through successive stages, the testee is administered a series of items which tend to alternate between items that are much too easy and items that are much too difficult. Reliability may be reduced by increased amounts of guessing toward the end of the test and by the possibility that such divergence, if perceived by the examinee, may have adverse psychological effects (see Weiss, 1974, p. 43).

Other research. A final study of flexilevel testing is currently in its preliminary stages. The objective of this study, as reported by Hansen, Johnson, Fagan, Tam and Dick (1974), is to explore the utility of adaptive testing procedures within the context of a computer-managed instructional system in an Air Force technical training environment. After extensive review of the literature on adaptive testing strategies, it was decided that the flexilevel model offered excellent potential for a 40-50% reduction in testing time along with either an increase in measurement accuracy or no decrease. In Phase I of this study (Hansen et al., 1974), a computer-administered flexilevel testing system was implemented. This flexilevel strategy differed from those used in previous studies in that examinees were individually entered into the flexilevel item structure based on estimates of their predicted performance derived from prior ability and performance data. After the administration of the flexilevel test, the remaining items in the structure were administered, yielding a total "conventional test" score. Thus, both flexilevel and conventional test scores were available for each individual. Although empirical data from this study have not yet been reported, Hansen et al. state that preliminary results support the feasibility and ease of implementing the flexilevel procedure and the capacity of the flexilevel testing strategy to offer considerable savings in testing time. However, firm conclusions regarding the results of this study must await the appearance of the results of empirical data analysis.

Summary. The studies to date of flexilevel testing have indicated that it can provide more accurate measurement than conventional tests for examinees whose ability levels differ from the average ability level of the group being tested, that scores on simulated flexilevel tests correlate highly with scores on the parent tests from which the former scores derive, and that the flexilevel test does increase the appropriateness of item difficulties for examinees' ability levels. On the other hand, results also indicate that the flexilevel test had lower internal consistency reliability and lower criterion-related validity than did conventional tests used for comparative purposes.

This conflicting series of results may be explained in part by the nature of the studies done. First, each study used a different research method; theoretical, real-data simulation, and actual test administration studies provide different kinds of information, and each type of study is subject to unique limitations. Second, the studies were all limited in the range of evaluative criteria used; only Olivier's (1974) study used more than one criterion of evaluation. Thus, there is little opportunity to compare results pertaining to one criterion of evaluation across two or more studies.

Objectives

Flexilevel testing strategies have not yet been evaluated in terms of such psychometric properties as the characteristics of the score distributions they yield, test-retest stability, parallel-forms reliability, correlations with direct criteria of ability, or precision of measurement when real item pools are used. The present series of studies was designed both to increase the extent and variety of information relevant to the comparison of flexilevel and conventional testing procedures and to attempt to clarify the interpretive difficulties raised by the results of the previous three studies.

To achieve these purposes, two related types of studies were done. First, flexilevel and conventional tests were computer-administered to college students. In view of the difficulties of paper and pencil administration found by Olivier (1974), computer-administration was felt to be better able to provide examinee response records containing no errors in branching and to eliminate the loss of records through such errors. Furthermore, since the computer can select the next item to be administered without the testee's knowledge of whether each item was answered correctly or incorrectly, computer administration might reduce somewhat the possible adverse psychological effects. The second study involved Monte Carlo simulation of examinee response records for the same flexilevel and conventional tests used in the computerized administration.

METHOD

Design

The empirical study, involving the actual computer-administration of flexilevel and conventional tests, was designed to permit the investigation of 1) the characteristics of the score distributions yielded by flexilevel and conventional tests; 2) the relationship between ability estimates yielded by the flexilevel and conventional tests; and 3) the test-retest stability of flexilevel and conventional test scores.

Because the generalizability of results yielded by an empirical study is frequently limited by the sample size and by the characteristics of the subjects tested, the procedures followed in the empirical study were also followed in a Monte Carlo simulation study. Monte Carlo simulation involves the generation of hypothetical groups of subjects and the use of either hypothetical or real item pools. The ability levels of the subjects and the item parameters are specified in advance. Then, using item characteristic curve theory and computer-generated random numbers, vectors of item responses are generated for a specified number of subjects. A study of this type provides no information on the psychological effects of testing on examinees and is limited by the assumptions used in generating response records for hypothetical testees, but it does provide large sample sizes and precise control of the characteristics of the population studied.

Thus, the Monte Carlo simulation study was designed to replicate the procedure followed in the empirical study and also to provide evaluative information beyond that provided by the empirical study. Paralleling the live-testing study, the simulation study provided information concerning score distributions and the relationship between scores on the flexilevel and conventional tests. Simulated re-administration of the same test, which under the conditions of empirical test administration provided test-ıetest stability data, provided data concerning the parallel-forms reliability of flexilevel and conventional tests. The availability of an ability criterion (i.e., knowledge of "true" scores) permitted the investigation of the relationships between ability estimates and underlying ability. Finally, since the items used in the simulation study were specified to have parameters identical to those items used in the empirical study, it was possible to replicate Lor 's (1971d) study of the amount of information or precision of measurement of each testing strategy using real, "non-ideal" items.

In summary, comparison of results of the two studies was considered to permit greater generality of conclusions than would be possible using only one method. Further, it was hoped that by following similar procedures in two different kinds of studies, sources of method variance leading to different conclusions could be identified.

Test Construction

Item Pool

The item pool used to construct the flexilevel 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 available for each item. Details concerning the development and norming of the item pool are reported by McBride and Weiss (1974). One characteristic of this item pool relevant to the evaluation of the flexilevel test was that there were many highly discriminating items of below average difficulty but considerably fewer highly discriminating and difficult items. Thus, in the item selection process, the more difficult items selected tended to be less discriminating than the less difficult items selected.

Flexilevel Test

Item structure. The flexilevel test constructed was one in which each examinee would attempt 40 items; thus, the total item structure required 79 items. These items were selected to be distributed along the difficulty continuum in the range of $b=-3.0$ to $b=+3.0$. Following Lord's (1971d) procedure, it was desired that the distance, d, between successive item difficulties be equal to a constant. Thus, the total range of difficulties divided by the number of intervals between 79 items (78) led to a desired value of d equal to .075. Of the available pool of items, only those with discrimination values greater than $a=.30$ were considered for inclusion in the flexilevel structure. The criterion for a constant distance between successive item difficulties was followed as closely as possible given the constraints of a real item pool and the minimum discrimination value required.

The mean difficulty of the 79 items in the flexilevel item structure was $b=-.01$; the mean discrimination value was $a=.65$, substantially greater than the minimum acceptable level. Table A-1 in the Appendix contains item reference numbers, item serial numbers, and difficulty and discrimination values for each item in the flexilevel structure. The item serial numbers, from 1 to 79, follow the rank order of item difficulties, from the least difficult, $b=-3.11$, to the most difficult, $b=2.95$, and are useful in determining the order in which items would be administered. Thus, the first item administered was always number 40 $(b=0.0)$; under the flexilevel branching rule a correct response would lead to the item whose serial number was the next larger one not previously administered, and an incorrect response would lead to the item whose serial number was the next smaller one not previously administered.

It may be noted that increasing item serial numbers do not always correspond to increases in b values (e.g., serial number 10 and 11).

This flexilevel test was constructed before the conclusion of the item norming studies which led to the publication of the characteristics of the final item pool (McBride & Weiss, 1974). Some of the item parameter estimates used in constructing the test were based on smaller sample sizes than those which characterized the final pool. Further norming studies led to some small changes in the "*b*" and "*a*" values characterizing certain items, and these changes did in some cases reverse the rank order of item difficulties. Fortunately, the changes were slight, and should not appreciably affect the adaptive property of the flexilevel test. The item parameters presented in Table A-1 are the final parameters, as reported in McBride & Weiss (1974).

Although the mean discrimination value of all 79 flexilevel test items was .65, the mean discrimination of the 40 items taken by any given examinee was a function of that examinee's ability, because of the relationship between item difficulty and item discriminating power. For example, an examinee who obtained 0 correct would have been administered items with mean *a*=.75, whereas an examinee obtaining 40 correct would have encountered items having a mean "*a*" value of .54. The mean "*a*" values corresponding to 10, 20, and 30 correct would be .74, .69 and .62, respectively. Thus, high ability examinees would be administered a less discriminating series of items than would low ability examinees.

Scoring. In the empirical study, the flexilevel test was scored using 1) simple number correct, and 2) Lord's (1971b) suggested modification in which an extra half-point is added to the score of each examinee responding incorrectly to the last item administered. This latter score was doubled, following Lord's suggestion, to eliminate the fractional values. Thus, the number-correct score, which will be referred to as Score 1, could range from 0 to 40. The half-point score, which will be referred to as Score 2, could range from 1 (the individual receiving ½ point, multiplied by 2, for an incorrect response to the final item) to 80 (all 40 items answered correctly).

In the simulation study, only Score 2 was calculated; this score uses more information than simple number correct and was also the scoring method used by Lord (1971d) in his theoretical studies. In addition, preliminary results from the empirical study suggested that the two scoring methods yielded essentially equivalent results.

Conventional Test

The conventional test, also consisting of 40 items, was the same test that was compared to two-stage testing procedures in studies by Betz & Weiss (1973, 1974). Item difficulties were concentrated around a "*b*" value of -.33 (somewhat easier than the median ability level of the group since guessing was a possibility). Again, a minimum *a* value of .30 was required; the resulting 40 items had a mean *a* of .54. Table A-2 in the Appendix provides the *b* and *a* values corresponding to each of the 40 items in the conventional test as reported by McBride & Weiss (1974). The test was scored using number correct, which ranged from 0 to 40.

Empirical Study

Administration and Subjects

Tests were administered to undergraduate and graduate students taking introductory psychology, introduction to statistics, and theory of measurement courses at the University of Minnesota in the fall of 1972. Students were tested at individual cathode-ray terminals (CRT's) connected by acoustical couplers to the University's CDC 6400 time-shared computer system (see DeWitt & Weiss, 1974, for details of the computer software system). Items were presented on the CRT screen, and testees indicated their response by typing in the number of the chosen alternative for each multiple-choice item. Following their response, the next item appeared on the screen. Instructional screens explaining the operation of the CRT's were provided prior to testing, and a proctor was present in the testing room to provide assistance to any testee having difficulty with the equipment.

Testees were permitted as much time as necessary to complete the tests and were so informed before test administration was begun. Testees received no feedback during the course of testing; at the end of the testing session they were told how many items they answered correctly.

Several subject groups were utilized in this study; these groups are summarized in Table 1. Subjects were administered two tests on each of two occasions. Reference to Table 1 shows that 477 subjects were tested

Table 1

Summary of Data Collection in the Empirical Study of Flexilevel Testing

	Time 1		Time 2	
Group	Tests Administered	N	Tests Administered	N
1 (Introductory Psychology)	Flexilevel and Conventional	107		
2a (Introductory Statistics)	Flexilevel and Two-stage	107	Flexilevel and Two-stage	94
2b (Introductory Statistics)	Two-stage and Conventional	110	Two-stage and Conventional	85*
3 (Theory of Measurement)	Flexilevel and Vocabulary norming	153	Flexilevel and Numeric norming	131
Total		477		310**

*Resulted in 74 usable conventional test-retest records
**Included 196 usable flexilevel test-retest records

on the Time 1 administration. Group 1 consisted of 107 students from the introductory psychology course; these students were administered the flexilevel and conventional tests. Group 2a consisted of 107 students from the introduction to statistics course; these students received a flexilevel test and a two-stage (see Weiss, 1974, pp. 3-11) test. Group 2b consisted of 110 students, also from the introduction to statistics course, who received the conventional and the two-stage test. Group 3 consisted of 153 graduate and undergraduate students from the theory of measurement course; these students received the flexilevel test and a series of difficult vocabulary items for use in continued norming of the vocabulary item pool.

Students from Group 2a were retested on the flexilevel and two-stage tests after an average interval of about five and one-half weeks. The students in Group 2b were retested on the conventional and two-stage tests, and those in Group 3 received a flexilevel retest and a series of number series items as part of the norming of an item pool to measure numeric problem-solving abilities. Students in Group 1 were not retested.

As Table 1 shows, the Group 1 data permitted the analysis of the relationship between scores obtained from flexilevel and conventional tests. Data from Groups 2a and 3 permitted the analysis of the test-retest stability of flexilevel test scores, and data from Group 2b permitted analysis of the stability of conventional test scores.

Table 1 indicates that retest records were not available for all of the students tested on the first occasion. Also, of the 225 students retested on the flexilevel test, only 196 of the test-retest records were usable, and of the 85 students retested on the conventional test, only 74 test-retest records were usable. This loss of examinee records was largely due to the failure of subjects to report for the retest. Computer failures during testing also contributed to incomplete and therefore unusable test records from both the Time 1 and Time 2 administrations.

Order effects. Since each student was administered two tests on each occasion, the possible effect of order of administration of the tests on obtained scores was a variable of interest, as it was in previous studies (e.g., Betz & Weiss, 1973; Larkin & Weiss, 1974; Larkin & Weiss, 1975). To study this variable, the order of administration in groups 1 and 2 was randomized so that approximately half of each group would receive the flexilevel test first (order 1), and the other half of the group would receive the flexilevel test second (order 2). The differences between mean scores from order 1 and order 2 were examined using t-tests for the significance of the difference between independent means.

Simulation Study

The Simulation Model

The Monte Carlo simulation procedure was initially developed for use in simulation studies of two-stage ability testing (Betz & Weiss, 1974); the procedure is described in detail in that report. The procedure was based on the assumptions and mathematics of item characteristic curve theory (Lord & Novick, 1968). The basic assumption made was that the probability of a correct response to an item is a generalized normal ogive function of an examinee's ability. To determine the probability of a correct response to an item given a specified ability level, the ability level and the normal ogive difficulty, discrimination, and guessing parameters corresponding to that item were entered into the equation suggested by Birnbaum (1968, Equation 17.3) and used by Lord (1971d,e) in his theoretical studies of flexilevel and two-stage testing.

The use of this simulation procedure in the study of two-stage testing (Betz & Weiss, 1974) yielded results which did not contradict and in most cases supported results obtained from a parallel empirical study (Betz & Weiss, 1973). Thus, it was considered to have utility for use in the present study.

Procedure

The computer program which "administered" the tests and calculated test scores in this study was a modification of the program used in the two-stage simulation study (see Betz & Weiss, 1974). The modification involved replacing the subroutine designed to administer a two-stage test with one that administered a flexilevel test. Following the design of the two-stage study, two administrations of the flexilevel test and two administrations of the conventional test were simulated for two samples of hypothetical testees.

One sample consisted of 10,000 testees sampled from a normally distributed population; ability levels were assigned to testees using a pseudo-random number generator which yielded a normally distributed set of numbers with mean 0 and variance 1. The second sample consisted of 1,600 testees, 100 at each of 16 ability levels between θ=-3.2 and θ=3.2. The 16 ability levels used are shown in Table 10. This latter distribution of ability levels, the "equal-frequency" distribution, was generated to allow calculation of values of the information function that were based on equal sample sizes at each selected point on the ability continuum.

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 the equation used to calculate the probability of a correct response to that item. Since the items were in a five-alternative multiple-choice format, the guessing parameter assigned to all items was .2, the probability of obtaining a correct response through random guessing. Following the calculation of the probability of a correct response, a random number was sampled from a rectangular distribution of real numbers between 0 and 1. If the random number was less than the former probability, the item was scored "1" (correct); if the random number was greater than the probability, the item was scored "0" (incorrect). The item response, 1 or 0, was then used in scoring the test and, in the flexilevel test, was used to determine the next item to be administered through the branching rules described previously.

Data Analysis

The following data were available from the empirical study: 1) conventional and flexilevel scores from Group 1; 2) test and retest score distributions for the flexilevel test from Group 2a and for the conventional test from Group 2b; and 3) flexilevel test and retest score distributions from Group 3.

One set of data from the simulation study consisted of ability level, scores from the two administrations of the flexilevel test, and scores from the two administrations of the conventional test for each of 10,000 "testees" whose ability levels were sampled from a normally distributed population.

The second set of simulation data consisted of ability level and the scores obtained from the two administrations of the flexilevel and conventional tests for 1600 "testees," 100 at each of 16 ability levels. These data were used only in the calculation of values of test information functions at each of the 16 ability levels.

Characteristics of Ability and Test Score Distributions

While it was assumed that the 10,000 ability levels sampled from a normally distributed population would be normally distributed, several characteristics of the resulting distribution of ability levels were examined to determine whether or not this assumption was reasonable. The mean, variance, skewness, and kurtosis for the 10,000 ability levels were calculated. These four statistics were then tested for the significance of their departure from expectation under the normality assumption (McNemar, 1969, pp. 25-28 and 87-88).

Analyses of the characteristics of the empirical and simulated test score distributions were done separately for each administration (test or retest) of the test. In the empirical study, analyses were also done separately for each subject group since the three groups were expected to differ in mean ability level.

Again, the mean, variance, skewness and kurtosis were calculated for each test score distribution; the indices of skewness and kurtosis were tested for the significance of their departure from normality. The flexilevel and conventional test score means within each group were compared using t-tests for the significance of the difference between the means of dependent groups (e.g., Glass & Stanley, 1970, pp. 297-300).

Reliability

Test-retest stability. Stability data for the flexilevel test were available from Groups 2a and 3 and for the conventional test from Group 2b.

Pearson product-moment correlations were calculated for the test-retest score distributions. To examine the effect of interval length on stability, the total groups were divided into three subgroups according to the length of the interval between test and retest. The three subgroups were: 1) short interval (13-30 days); 2) moderate interval (31-46 days); and 3) long interval (47-62 days).

These intervals, determined so that the three subgroups would be of approximately equal size, were the same intervals used in the study of pyramidal adaptive testing procedures (Larkin & Weiss, 1974). Product-moment correlations were calculated between test and retest scores within each subgroup.

In addition to the possibility that test-retest stability might be affected by interval length was the possibility that the stability of flexilevel and conventional tests might be differentially affected by memory of particular items and of the previous responses to them. To the extent that memory leads the examinee to repeat the same responses he/she made before, the similarity of results on two test administrations tends to be increased. This inflation of the stability coefficient can logically be assumed to be directly related to the number of items repeated on the retest.

In re-administering the conventional test, all 40 items were repeated. However, the number of items repeated in an adaptive retest varies with the adaptive strategy and with the particular individual's response patterns. The number of items repeated in a 40-item flexilevel test can range from 1 to 40. In order to assess the magnitude of memory effects on the stabilities of the flexilevel and conventional tests, a distribution of the number of items repeated on the flexilevel retest was obtained. The number of items repeated in a flexilevel test is equal to the number of items in the test minus the difference between the number-correct scores obtained by an examinee on test and retest. The relationship between the number of repeated items and the size of the stability coefficient was examined.

Parallel forms reliability. In the two simulated administrations of each test, examinee ability level and the parameters assigned to each item were constant, thus yielding the same probability of a correct response for any given item-individual interaction. However, the random number determining the scoring of the given item varies so that simulated re-administration of the same test may yield a different pattern of right and wrong answers and, in the case of the flexilevel test, differences in the branching pattern. Thus, simulated re-administration of the same test can be used to evaluate parallel-forms reliability; while the item parameters are identical between the two forms, there is no specific item content overlap since only item *parameters* determine response patterns in a simulation study.

The operation of the simulation computer program was such that each "run" of the program provided ability levels and test scores for 100 "examinees." For each group of 100, Pearson product-moment correlation coefficients were calculated to express the degree of relationship between scores obtained from the two simulated administrations of each test. Thus, there were 100 reliability coefficients for each test obtained from 100 samples from a hypothetical population with a normal distribution of underlying ability. The mean and standard deviation of the obtained sampling distributions were used to construct confidence intervals indicating the effective range of reliability coefficients obtained in replications of the study using samples of 100. The 95% confidence intervals were obtained by adding to and subtracting from the mean the value of two standard deviations of the obtained sampling distribution. In addition, taking the mean of each sampling distribution as an estimate of the population reliability (ρ), the standard errors of the mean were calculated and used to test the significance of the difference between the expected reliability values for the conventional and flexilevel tests in the population.

In a previous study (Betz & Weiss, 1974) the product-moment coefficients were transformed to Fisher's Z_r values so that the effects of possible non-normality in the original distribution of *r* coefficients on the length and symmetry of the confidence intervals around the expected value could be evaluated. However, the expected values and confidence intervals obtained using the normalized *r* values were found to be identical to those obtained from the original distribution, and it was concluded that skewness in the latter distribution was not a factor influencing the obtained confidence intervals. Since the parallel-forms correlations in the present study were expected to be similar in magnitude to those of the previous study, only the original distribution of *r* values was used to derive expected values and confidence intervals.

Relationships Between Flexilevel and Conventional Test Scores

The examinees in Group 1 were administered both the flexilevel and conventional tests. To analyze the relationship between the flexilevel and conventional 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. Similar analyses were completed for the simulated distributions of 10,000 flexilevel and conventional test scores.

Relationships Between Test Scores and Underlying Ability

Product-moment and eta coefficients were calculated to determine the nature and degree of relationship between each set of 10,000 scores and the distribution of underlying ability. In addition, the characteristics of the sampling distribution of 100 *r* values obtained from the 100 samples of 100 "subjects" were evaluated; confidence intervals indicating the effective range of values were constructed and tests were made of the significance of the difference between the means of the obtained sampling distributions.

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 accuracy 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):

$$I(\Theta) = \left[\frac{\frac{\partial}{\partial\Theta}\,\varepsilon(x|\Theta)}{\sigma_{x|\Theta}}\right]^2. \qquad [1]$$

where $I(\theta)$ indicates the amount of information provided by a given test, scored in a specific way, at a given level of underlying ability θ. The numerator in Equation 1 is the slope of the regression of observed test scores on underlying ability (calculated by evaluating the first derivative of the regression function at 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(θ).

According to Lord (1970), the numerator of Equation 1 represents the capability of test scores to differentiate among examinees with ability levels in the immediate vicinity of θ. 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

$$\frac{x_2-x_1}{\theta_2-\theta_1} \qquad [2]$$

indicates the degree to which the test discriminates these two ability levels.

The denominator of Equation 1 is the conditional standard error of measurement at a particular level of ability. The square root of I(θ) is inversely related to the confidence interval for estimating observed score from underlying ability (Green, 1970). Thus, a low value of I(θ) indicates a larger standard error of measurement at a particular level of ability, and the higher the value of I(θ), the smaller the error of measurement.

The procedures used to calculate the relative amount of information provided by the flexilevel and conventional tests for both the normal and "equal-frequency" distributions of ability were identical to those used in the earlier simulation study of two-stage testing (Betz & Weiss, 1974). 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 program. The third degree 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 first derivative of the third degree polynomial was then derived so that the slope of the regression function could be calculated at the desired θ levels.

The normal ability distribution was divided into 33 intervals between θ=-3.3 to θ=+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 θ=-3.3 to θ=-3.1, and θ=-3.2 was taken as its midpoint. For each interval, the variance of the test scores of individuals whose generated 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.

In order to obtain information values with more equivalent stability across the ability continuum, the "equal-frequency" distribution with 100 "examinees" at each of the 16 ability levels shown in Table 10 was used. While the numerator of Equation 1 used slope values based on the first derivative of the regression equation derived from the normally distributed population (thus yielding slope values based on different sample sizes), the $\sigma_{x|\theta}$ values in the denominator of Equation 1 were all calculated using samples of 100. Thus, in the "equal-frequency" distribution of ability, the numerator of the information equation was the slope at one of the 16 ability levels, and the denominator was the standard deviation of the 100 scores generated at that level.

RESULTS

Order Effects

Table 2 presents the results of the analysis of the effects of order of administration on the means of the obtained test scores. Results are indicated for both methods of scoring the flexilevel test (Lord, 1971b). As the table indicates, there were no significant differences in mean scores as a function of order of administration for either of the groups. These results correspond to previous findings that order of administration does not affect scores on conventional tests (Betz & Weiss, 1973; Larkin & Weiss, 1974), two-stage tests (Betz & Weiss, 1973), pyramidal tests (Larkin & Weiss, 1974), or two adaptive tests taken in combination (Larkin & Weiss, 1975).

Table 2

Flexilevel Test Score Means and Standard Deviations for Subgroups Completing the Flexilevel Test in Different Orders

Group and Score	Order 1: Flexilevel First			Order 2: Flexilevel Second			Test of Significance		
	N	Mean	S.D.	N	Mean	S.D.	t	df	p
Group 1									
Score 1	54	19.37	6.09	53	19.34	4.99	.03	105	.97
Score 2	54	39.17	12.00	53	39.17	10.00	.00	105	.99
Group 2a									
Score 1	57	22.35	4.75	50	21.92	6.22	.41	105	.68
Score 2	57	45.19	9.43	50	44.20	12.36	.47	105	.63

Ability and Test Score Distributions

Empirical study. Table 3 contains data describing the flexilevel and conventional test score distributions; results are presented separately for each subject group since the groups were expected to differ in mean ability level.

Table 3 shows that the mean Score 1 on the first testing with the flexilevel test was lowest for Group 1 (19.36), next higher for Group 2a (22.15), and highest for Group 3 (27.08). The differences between each of these group means were statistically significant (p<.01). These results indicate significant differences in ability level in the three groups and were expected since Group 1 consisted of beginning undergraduate students, Group 2a consisted of somewhat more advanced undergraduates, most of them psychology majors, and Group 3 consisted of honors undergraduate and graduate students. The standard deviations of scores in the three groups indicated essentially equivalent within-group variability among the groups.

Differences among the three groups were also reflected by the skewness of the score distributions. The group 1 scores were significantly positively skewed, indicating a concentration of lower scores, while those of Group 3 were significantly negatively skewed, indicating a predominance of higher scores. The scores of Group 2a were not skewed. The Group 2a score distribution was somewhat platykurtic, indicating a more even spread of scores than was the case in Group 1, where scores were normally peaked, or Group 3, in which the scores tended to be more peaked than is typical of a normal distribution.

Group differences were also reflecte' by mean scores on the conventional test. The mean score obtained by Group 1 (18.58) was significantly (p<.01) less than the mean score obtained by Group 2b (24.19). The variability of the scores in the two groups was almost identical (8.22 and 8.28). Further, the Group 1 conventional test scores were again significantly positively skewed, while those of Group 2b were not skewed. The Group 2b scores were also significantly platykurtic on Time 1, indicating that the distribution of scores was flatter than a normal distribution of scores.

The shape of the score distributions indicates that the difficulty levels of both the flexilevel and conventional tests were most appropriate for the individuals sampled from the Group 2 population. However, an analysis of the mean number-correct scores in relation to expected means offers further information concerning the appropriateness of t^h tests for measuring groups of individuals differing in ability level.

The mean difficulty of the conventional test items was $b=-.33$, corresponding to a p (proportion correct) value of .57 in the norming sample. This p value should result in a mean number correct of 23 of the 40 items administered in samples of examinees similar in ability to the norming group. In Group 2b, the mean number correct on the conventional test was 24.19, close to that expected, while the mean number correct in Group 1, 18.58, indicates that the conventional test items were somewhat too difficult for the group as a whole.

Table 3

Descriptive Data for Flexilevel and Conventional Test Score Distributions for Initial Test (Time 1) and Retest (Time 2), by Subject Group

Group and Test	N		Mean		Standard Deviation		Skew		Kurtosis	
	Time 1	Time 2	Time 1	Time 2	Time 1	Time 2	Time 1	Time 2	Time 1	Time 2
Group 1										
Flexilevel										
Score 1	107		19.36		5.55		.47*		.02	
Score 2	107		39.17		10.99		.48*		.00	
Conventional	103		18.58		8.22		.58*		-.48	
Group 2a										
Flexilevel										
Score 1	107	94	22.15	23.09	5.47	5.27	.08	-.05	-.55	-.69
Score 2	107	94	44.73	46.63	10.85	10.54	.04	-.07	-.55	-.70
Group 2b										
Conventional	110	85	24.19	25.67	8.28	8.32	-.04	-.24	-1.01**	-.93
Group 3										
Flexilevel										
Score 1	153	131	27.08	27.63	6.17	5.76	-.61*	-.45*	.79	.27
Score 2	153	131	54.80	55.61	11.88	11.36	-.55	-.45*	.71	.25

* Skew index significantly different from zero (p<.05)

** Kurtosis index significantly different from zero (p<.05)

For the flexilevel test, the mean item difficulty was $b=-.01$, yielding a p value of .50 or an expectation of 20 correct of the 40 items administered. However, the flexilevel test is also designed so that the items administered to a given examinee are more appropriate to his/her ability level, or in other words, closer to $p=.50$ difficulty for the examinee. Thus, while mean score differences on a flexilevel test should to some extent reflect group differences in ability, they should also tend to be closer to 50% correct in different subject groups than should mean scores on a conventional test.

Comparing the flexilevel and conventional test score means within groups indicates that for Group 1, the mean number correct (flexilevel mean Score 1 equal to 19.36 and conventional score mean equal to 18.58) was not significantly different for the two groups; however, the flexilevel mean was closer to the expectation of 50% or 20 items correct than was the conventional test to the expectation of 57% or 23 items correct. In Group 2, the mean conventional test score (24.19) was significantly ($p<.01$) greater than the mean flexilevel Score 1 (22.15). The conventional test mean was close to its expectation, but the flexilevel mean was again closer to 50% correct. If item difficulty were the only factor influencing the mean scores, a higher conventional test mean would be expected in both groups since these items were somewhat easier, on the average, than the items in the flexilevel test. Thus, it appears that the flexilevel test does adapt item difficulties to the ability levels of individuals within groups and across groups differing in ability. The fact that the flexilevel test score means were closer to .50 also implies less guessing on the flexilevel test.

Further comparison of flexilevel and conventional test score distributions indicates that for Group 1, conventional test scores were significantly more variable ($p<.01$) than were the flexilevel scores. Both distributions were significantly positively skewed, reflecting the lower ability level of Group 1 as a whole. The conventional score distribution showed a non-significant tendency toward flatness (platykurtosis) not shown by the flexilevel scores.

In Group 2, the conventional test scores were again significantly more variable. Neither distribution of scores was skewed, although both tended to be flatter than a normal distribution; the latter tendency was statistically significant only for the conventional test.

<u>Simulation study</u>. The assumption that 10,000 ability levels sampled from a normally distributed population would themselves be normally distributed was accepted. The mean ability level was 0.0, the variance was 1.0, and the degrees of skewness and kurtosis of the ability distribution did not show significant departures from normality.

Table 4 presents data describing the distributions of 10,000 scores generated in the simulation study. The data for the flexilevel test may be compared to that of Score 2 in the empirical study, as shown in Table 3.

Table 4

Descriptive Data for Flexilevel and Conventional Test Score Distributions Generated by Monte Carlo Simulation with an Underlying Normal Distribution of Ability

Test	N	Mean	S.D.	Skew	Kurtosis
Flexilevel (Score 2)					
Time 1	10,000	46.6	10.06	-.23*	-.29*
Time 2	10,000	46.6	10.08	-.24*	-.31*
Conventional					
Time 1	10,000	25.9	6.48	-.25*	-.46*
Time 2	10,000	25.9	6.43	-.23*	-.53*

*Statistically significant at p<.01

The flexilevel Score 2 mean was 46.6, corresponding closely to that obtained by group 2a in the empirical study (44.73). This Score 2 mean indicates that the mean number correct (Score 1 in the empirical study) for the simulated examinees was about 23. The conventional test score mean (25.9) was most similar to that obtained by Group 2b in the empirical study (24.19). The agreement of the simulated data with that of Group 2 in the empirical study was expected; other samples from the Group 2 population (introductory statistics students) comprised a large proportion of the original item norming samples, and the average ability level of this group was at about the mean of the norming population as a whole.

The mean number correct on the conventional test (25.9) was significantly greater (p<.01) than the mean number correct on the flexilevel test (assuming it to be 23); this result is again in agreement with that found for Group 2 in the empirical study. The standard deviation of the flexilevel number correct scores was about 5.0 (since the variability of Score 1 was shown in the empirical study to be roughly half that of Score 2) as compared to a standard deviation of about 6.5 for the conventional test scores. While the conventional scores were again more variable than the flexilevel scores, score variability for both tests in the simulation study was uniformly lower than that shown in the empirical study.

Both the flexilevel and conventional test score distributions were significantly negatively skewed and significantly platykurtic. However, the flexilevel scores were less platykurtic than were the conventional scores, indicating that the former more closely reflect the known underlying normal distribution of ability. The direction of skewness for the flexilevel scores paralleled the negative skew found for Group 3 in the empirical study, although the absolute degree of skewness was less in the simulation study. The skewness of the conventional test scores was closest in degree to that found in the Time 2 administration in Group 2b in the empirical study. The platykurtosis characterizing both simulated score distributions is in agreement with that shown by Group 2 in the empirical study, although not Groups 1 and 3.

Reliability

Test-retest stability. Table 5 contains the test-retest stability correlations for the flexilevel and conventional tests, as obtained from the empirical study. The first set of columns indicates the stability of each test for the total group of examinees; the last three sets of columns show stability as a function of the length of the interval between test and retest.

Table 5

Test-retest Stability Correlations as a Function of Interval Length, and for Total Group (empirical data)

			Retest Interval (in days)					
	Total Group		13-30		31-46		47-67	
Test	N	r	N	r	N	r	N	r
Flexilevel--								
Score 1	194	.89	53	.92	91	.86	50	.88
Score 2	194	.89	53	.93	91	.86	50	.87
Conventional	74	.89	25	.89	28	.91	21	.87

The overall stability of scores on the two tests was comparable; both had test-retest correlations of .89. Stability was not clearly related to interval length for either testing strategy. Scores on the flexilevel test were most stable over the shortest interval (r=.92 or .93), but stability over the two longer intervals was about the same. In contrast, scores on the conventional tests were most stable over the moderate interval (r=.91), and least stable over the longest interval (r=.87). The flexilevel test scores were more stable over a short time interval than were scores on the conventional test; conventional test scores were more stable in the moderate time interval; and scores on the two testing strategies showed equal stability in the long time interval.

Table 6 indicates the number of examinees repeating 40, 39, 38 or 37 or fewer items on the retest, the mean and standard deviation of the number correct (Score 1) obtained by each group of examinees, and the stability of scores within each group.

Table 6

Stability of Flexilevel Test Scores (Score 1) as a Function of the Number of Items Repeated

Number of Items Repeated	N	Mean		Standard Deviation		r
		Time 1	Time 2	Time 1	Time 2	
40	39	27.13	27.13	5.13	5.13	1.00
39	63	25.03	25.33	6.55	6.33	.99
38	40	26.42	26.82	5.82	5.98	.94
37 or less	52	22.87	25.24	6.60	6.28	.64

As shown in Table 6, almost three-fourths of the total group repeated 38 or more items; only 52 of the 194 examinees repeated 37 or less.

Examinees who repeated 40 items obtained the same test score on both administrations of the test; thus, there must of necessity be no change in the mean score from Time 1 to Time 2 and a correlation of 1.0 between the two sets of scores. For examinees who repeated 38 or 39 items, mean scores showed an insignificant increase from Time 1 to Time 2. The stability of scores within these groups, r=.99 and r=.94, while probably partly an artifact of the only 1 or 2 point score changes shown from Time 1 to Time 2, also indicates high consistency in the direction of score changes in terms of maintaining at Time 2 the rank order established on the Time 1 administration.

The performance of examinees repeating fewer than 38 items was markedly different from that of the other examinees in several respects. The mean score on Time 1 for this group was lower than that for the other three and was significantly (p<.01) poorer than was the performance of examinees repeating 38 or 40 items. However, this group of examinees showed a significant (p<.01) increase in the mean score obtained on the Time 2 administration; this increase (from 22.87 to 25.24) brought the performance of the group repeating fewer than 38 items to a level comparable to that of the groups repeating 38 or more. Finally, test-retest stability dropped markedly in this group, from r=.94 in the "38" group to r=.64 in the "37 or less" group.

From these results it would appear that the overall stability of the flexilevel test (r=.89) reflects the combined effects of 1) a majority of examinees whose performance from Time 1 to Time 2 was highly stable in terms of both rank order and overall level of performance, and 2) a small group of examinees whose overall performance was initially at a significantly lower level than that of the larger group, whose mean score increased significantly on the Time 2 testing, but who showed far less consistency in rank ordering from Time 1 to Time 2.

Parallel forms reliability. Table 7 presents the characteristics of the sampling distributions of parallel forms reliability coefficients obtained from the simulation study. These data show that the flexilevel test was more reliable than the conventional test, having a mean reliability of .84 as contrasted with that of .80 for the conventional test. This difference was statistically significant at p<.001.

Table 7

Characteristics of Sampling Distributions of Parallel Forms Reliability Coefficients Using 100 Random Samples of 100 "Testees"

			Range		95% Confidence Interval (±2 S.D.'s)	
	Mean	S.D.	Maximum	Minimum	Upper	Lower
Flexilevel	.84	.029	.90	.74	.90	.78
Conventional	.80	.038	.88	.65	.87	.72

The standard deviation and range of coefficients obtained for the flexilevel test was also smaller than the values for the conventional test, indicating more consistency in the reliability estimates obtained from the 100 samples. The obtained 95% confidence intervals indicate that the effective range of reliability coefficients based on sample sizes of 100 for the flexilevel test was between .78 and .90, while that for the conventional test was between .72 and .87.

Relationships between Flexilevel and Conventional Test Scores

Table 8 presents the product-moment correlations and eta coefficients describing the relationship between flexilevel and conventional test scores for both the empirical and simulation data. All of the obtained coefficients were significantly different from zero ($p<.001$), and none of the eta coefficients indicated a significant degree of non-linearity in the relationship between the two distributions of scores.

Table 8

Relationships between Flexilevel Scores (Score 2) and Conventional Test Scores

	Time 1
Empirical Study (Group 1, N=103)[a]	
Product-moment correlation	.89
Regression of flexilevel scores on conventional scores (eta)	.90
Regression of conventional scores on flexilevel scores (eta)	.91
Simulation Study (Time 1, N=10,000)[b]	
Product-moment correlation	.82
Regression of flexilevel scores on conventional scores (eta)	.82
Regression of conventional scores on flexilevel scores (eta)	.82

[a] Four subjects were eliminated from this analysis because of incomplete response records on either the flexilevel or conventional test

[b] Data for Time 2 are not shown since the results were identical to the time 1 data

The relationship between scores was higher in the empirical study than in the simulation study; in the former, r=.89 with eta coefficients of .90 and .91, and in the latter both r and eta coefficients were equal to .82. Thus, flexilevel test scores accounted for about 81% of the variance in conventional test scores in the empirical study, but for only about 67% in the simulation study.

Relationships between Test Scores and Ability

The product-moment r and eta coefficients summarizing the extent of relationship between flexilevel test scores and generated underlying ability ("validity") in the simulation data were equal to .91 for both flexilevel "administrations," as calculated using all 10,000 scores. The coefficients for the conventional test and ability were both equal to .89. Both sets of coefficients indicated a high linear relationship between test scores and ability, although the flexilevel test showed a significantly ($p<.001$) higher relationship. Thus, underlying ability level accounted for approximately 83% of the variance in flexilevel test scores and for approximately 79% of the variance in conventional test scores.

Table 9

Characteristics of Sampling Distributions of Product-moment Correlations between Test Scores and Simulated Ability Calculated on 100 Samples of 100 Subjects

			Range		95% Confidence Interval (±2 S.D.'s)	
Variables	Mean	S.D.	Maximum	Minimum	Upper	Lower
Flexilevel--Ability						
Time 1	.91	.015	.95	.87	.94	.88
Time 2	.91	.015	.95	.87	.94	.88
Conventional--Ability						
Time 1	.89	.020	.93	.81	.93	.85
Time 2	.89	.019	.93	.85	.93	.85

Table 9 presents the characteristics of the sampling distribution of product-moment coefficients calculated on 100 groups of 100 testees. A comparison of the mean values shown in Table 9 with those calculated for the total distribution of 10,000 sets of scores (.91 for flexilevel, .89 for conventional) shows that the two methods gave identical results: the mean r for flexilevel was .91, and the mean r for the conventional was .89.

Examination of the confidence intervals shows that, for flexilevel, the effective range of correlations with ability over 100 samples was between .88 and .94, while that for the conventional test was between .85 and .93. The difference between the means of the obtained sampling distributions was statistically significant at p<.001.

Information Functions

Equal-frequency distribution. Table 10 presents estimated values of the information function (I(θ)) for the flexilevel and conventional tests at each of sixteen ability levels. The values at each level were obtained through application of the method of "moving averages" (McNemar, 1969, p. 8) to the average of the values obtained from the two administrations of each test. Thus, the values in Table 10 represent "best" average estimates of the value of the information at each ability level. Table A-3 in the Appendix indicates the information values averaged over the two administrations (but before application of the method of "moving averages"), separate values for the first and second administrations of each test, and the mean and standard deviation of information values over the 16 ability levels used.

Table 10

Values of the information function (I(θ)) for flexilevel and conventional tests at points along the continuum of underlying ability (equal-frequency distribution)

Level of Ability (θ)	Flexilevel	Conventional
3.2	.18	.32
3.0	.66	.82
2.5	1.71	1.71
2.0	3.20	2.84
1.5	4.72	3.93
1.0	5.76	4.53
.5	6.38	4.76
.1	6.62	4.65
-.1	6.70	4.41
-.5	6.38	4.04
-1.0	5.80	3.59
-1.5	4.81	2.99
-2.0	3.88	2.25
-2.5	2.90	1.44
-3.0	2.10	.74
-3.2	1.13	.27
Mean	3.86	2.68
S.D.	2.23	1.62

Note. Values obtained using method of "moving averages" (McNemar, 1969, p. 8)

The data contained in Table 10 are summarized in graphic form in Figure 4.

The shape of the information curve for the conventional test, as shown in Figure 4, is very similar to that found in Lord's (1971d) theoretical study; that is, the information values are highest near the center of the ability distribution and drop off sharply at the extremes. 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 construction of a perfectly peaked conventional test, both 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 4

INFORMATION FUNCTIONS FOR FLEXILEVEL AND CONVENTIONAL TESTS USING "EQUAL-FREQUENCY" DISTRIBUTION OF ABILITY

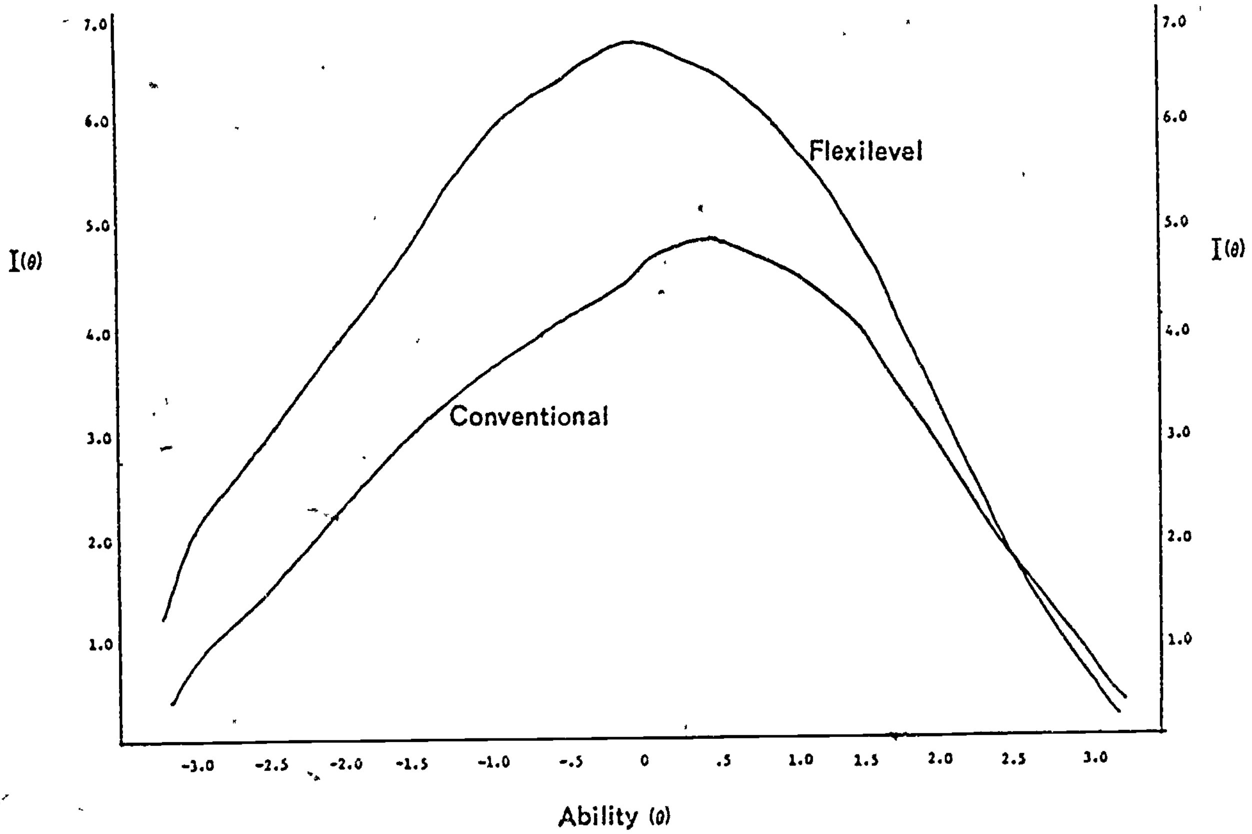


Figure 4 also shows that the flexilevel test, while providing more information than the conventional test for ability levels between θ=-3.2 and θ=2.0, did *not* provide more constant accuracy of measurement across all ability levels. Contrary to Lord's results, in which the flexilevel test showed a more nearly horizontal information function, the shapes of the two information functions shown in Figure 4 are actually quite similar; both tests showed greatest accuracy near the ability level corresponding to the mean difficulty of the test items, and a substantial drop in accuracy at more extreme ability levels.

The overall level and shape of the information functions shown in Figure 4 are also reflected by the means and variances of the information values for each test, as shown in Table 10. The mean value for flexilevel (3.86) was higher than that for the conventional test (2.68), but the standard deviation of information values for the flexilevel test (2.23) was greater than that for the conventional test (1.62). The larger standard deviation for the flexilevel test reflected a greater degree of variation in information values across the sixteen levels of ability.

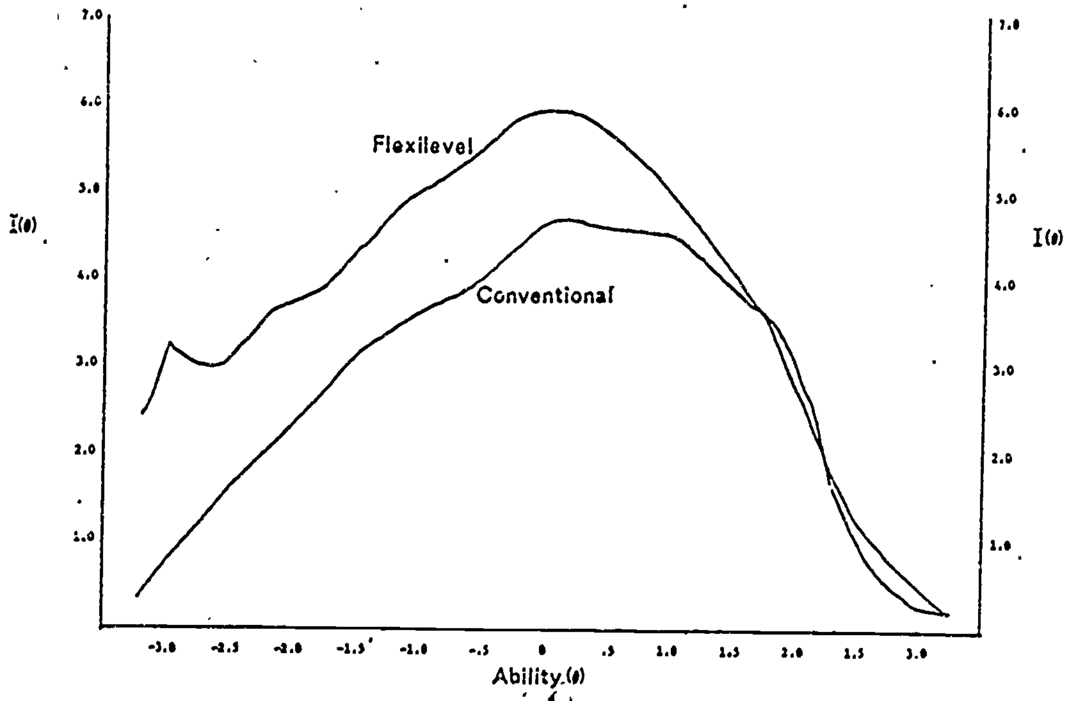
The data in Table 10, representing "best" estimates of the value of information at each ability level may be compared with that in Appendix Table A-3, in which the Time 1 and Time 2 results are presented separately. The data in Table A-3 indicate that while the means and standard deviations of information values were similar for the two test administrations, there were substantial differences in the information values at a given ability level. For example, at θ=2.0, the Time 1 administration resulted in a flexilevel information value of 3.47, while the Time 2 value for flexilevel was 2.33. For the Time 1 administration, the flexilevel test provided most information (7.18) at θ=-1.0, while the greatest amount of information in the Time 2 administration was provided at θ=-.1. Similar differences due to sampling error were found for the conventional test.

Normal distribution. Appendix Table A-4 presents the estimated values of I(θ) provided by the flexilevel and conventional tests when calculated using subjects with an underlying normal distribution of ability; again, these values were obtained by application of the method of "moving averages" to the averages of the Time 1 and Time 2 administrations. Table A-5 in the Appendix contains the initial average information values, the separate values for the first and second test administrations, the mean and standard deviation of the 33 values for each test and the number of "testees" assigned ability levels within each interval of ability.

The results indicated in Table A-4 are summarized graphically in Figure 5.

Figure 5

INFORMATION FUNCTIONS OF FLEXILEVEL AND CONVENTIONAL TESTS USING NORMAL DISTRIBUTION OF ABILITY



As shown in Figure 5, both the flexilevel and conventional tests again show greatest accuracy of measurement at the ability level corresponding to the mean difficulty of the items, and losses of accuracy at the extremes. Again, flexilevel provides more information between θ=-3.2 and about θ=1.5, but at ability levels greater than θ=1.5, the two tests yield essentially equal information values.

The means and standard deviations of the information values, as shown in Table A-4, indicate that the flexilevel test provided a higher overall level of information (3.81) than did the conventional test (2.85) but that its information values were also slightly more variable (1.70 to 1.53). Again, the results obtained from the separate administrations of each test (as shown in Appendix Table A-5) indicate substantial variability in the information values corresponding to each ability level.

DISCUSSION

Comparison of the score distributions obtained from three groups of subjects in the live testing indicated that both the flexilevel and conventional tests reflected differences in the mean ability levels of the three subject groups in terms of both the mean number-correct obtained by each group and the skewness of the group score distributions. In terms of these two characteristics, it appeared that the average difficulty level of the test items was most appropriate for Group 2. In this group, both score distributions tended to be platykurtic, although the degree of platykurtosis was greater for the conventional test and was statistically significant on the Time 1 administration. These findings are in agreement with previous findings (Betz & Weiss, 1973; Larkin & Weiss, 1974) showing that conventional tests yielded score distributions that were more platykurtic than the adaptive (two-stage and pyramidal) tests with which they were compared. In the present study, conventional test scores were more variable than the flexilevel scores.

While the flexilevel test did reflect differences in the ability levels of the groups, it was also found to adapt item difficulties to differences in the ability levels of examinees within groups; this was inferred from the fact that in Groups 1 and 2, the mean number-correct for the flexilevel test was closer to 50% correct than it was for the conventional test, even though the mean difficulty level of the items in the two tests would have implied otherwise. This finding is similar to that found by Larkin & Weiss (1974) for pyramidal adaptive tests, in which the mean number of items answered correctly was slightly more than half of the 15 items administered. These results suggest that adaptive tests reduce random guessing, since the mean number correct was close to that expected from free-response items, although the test used multiple-choice items.

The score distributions yielded in the simulation study were most similar to those yielded by Group 2 in the empirical study in terms of the mean number-correct and the tendency toward platykurtosis. The simulated score distributions for both tests were significantly negatively skewed and significantly platykurtic, but the flexilevel test better reflected the underlying normal distribution of ability. Again, conventional scores were more variable than flexilevel scores, but both sets of scores were uniformly less variable than those in the empirical study. In the simulation study of Betz & Weiss (1974), two-stage tests were also found to better reflect the underlying normal distribution of ability than did the conventional test, but, again, all score distributions were significantly platykurtic, and the score distributions of the conventional test and one of the two-stage tests were significantly negatively skewed.

Comparing the empirical and simulation studies indicated that, as in Betz & Weiss (1974), real testees obtain scores that are uniformly more variable than are scores generated in the simulation studies. In contrast to simulated examinees, actual testees differ from each other on variables in addition to ability level. Differences in motivation to do well, anxiety level, and tendency to guess may contribute to additional variance in test scores obtained from live test administration.

The test-retest stability of scores from both tests was identical; both had stability coefficients of r=.89. No consistent relationship between stability and the length of the interval between test and retest was found for either test, although the flexilevel test was more reliable in the short time interval. The stability of flexilevel test scores (r=.89) was identical to that found for scores from a 40-item two-stage test (Betz & Weiss, 1973). The stability of scores on a 15-item pyramidal test was found to range between r=.79 and r=.89 for different methods of scoring the test; the modal correlation was r=.86 (Larkin & Weiss, 1974). These data suggest that the pyramidal testing strategy, which with 15 items achieved stabilities as high as the 40-item flexilevel test, is a more efficient method of adaptive testing.

The analysis of the possible effects of memory of items repeated on the size of stability coefficients calculated in the live-subject group showed that on the flexilevel test, three-fourths of the total group repeated 38 or more of the 40 items administered. Since the number of items repeated in the flexilevel test could vary between 1 (the first item administered to all examinees) and 40, the fact that most people repeated 38 to 40 items indicates substantial consistency in the responses of examinees over the two test administrations. This in turn would appear to imply that the flexilevel tailors item difficulties to be appropriate to each examinee's ability, for example, low ability examinees receiving many items that are too difficult for them would be likely to perform inconsistently over two test administrations because of the possible effects of random guessing.

Further, the stability of scores for examinees who repeated 38 to 40 items on the flexilevel test was higher (r=.94 to r=1.00) than the stability of conventional test scores, on which examinees repeated all 40 items (r=.89). Thus, when the two tests were roughly equated for the effects of memory, the flexilevel test yielded more stable scores. This finding is in agreement with the findings of Betz & Weiss (1973) in which scores from a two-stage test were more stable than those from a conventional test when the effects of memory were equated, and the findings of Larkin & Weiss (1974) which showed that memory was operating to inflate the stability of conventional test scores.

The flexilevel test had significantly higher parallel forms reliability (r=.84) than did the conventional test (r=.80), as determined from the simulation study. The reliability of the flexilevel test compares favorably to parallel forms reliability coefficients of r=.76 and r=.83 for two two-stage tests as found in the simulation study of Betz & Weiss (1974); in that study, the reliability of the conventional test was also r=.80.

In both the present simulation study and that of Betz & Weiss (1974), however, there was substantial variability among the reliability coefficients calculated across 100 samples of size 100. In the present study, the effective range of coefficients (..e., 95% of those obtained) was between .78 and .90 for the flexilevel test and between .72 and .87 for the conventional test. This finding has implications for the interpretation of results of simulation studies based on single samples of 100 or fewer "subjects" (e.g., Jensema, 1972; Urry, 1970); in such cases, obtained reliability or validity coefficients may not be representative of results that would be obtained over a larger number of samples or using a single large sample.

Parallel forms reliability as determined from the simulation study was expected to be lower than the test-retest stability because it includes as systematic score variance fewer kinds of specific or error variance (Stanley, 1971). A test-retest stability coefficient includes as systematic variance two sources of variance which are treated as error in a parallel-forms design: 1) variance specific to the content of particular items, and 2) actual memory of particular items and of the previous responses to them. Thus, when the factors of item content sampling and memory are significant sources of variance, test-retest stability coefficients will be higher than parallel-forms coefficients. It may be noted that there was a larger difference between stability and parallel-forms reliability for the conventional test (r=.89 versus r=.80) than there was for the flexilevel test (r=.89 versus r=.84). Since there is no reason to suspect differences between the two tests in content-specific variance, the greater difference for the conventional test supports a hypothesis that the stability for the conventional test is inflated more by memory factors.

The correlation between flexilevel and conventional test scores obtained from the same sample of examinees in the empirical study was .89, indicating that the two sets of test scores share about 80% common variance. In contrast, the correlation found in the simulation study was only .82, indicating about 67% shared variance. This difference between empirical and simulated data was not found in the studies of Betz & Weiss (1973, 1974), in which the correlations between conventional and two-stage tests ranged between .79 and .84 in both the empirical and simulation studies. Further, in other empirical studies, correlations averaging .84 were found between conventional and pyramidal tests (Larkin & Weiss, 1974), and correlations between .79 and .84 were found between two-stage and pyramidal tests (Larkin & Weiss, 1975). Thus, the correlation of .89 between flexilevel and conventional test scores found in the present empirical study is higher than those found in the parallel simulation study or in other studies comparing two or more testing strategies.

The flexilevel test had a significantly higher relationship to underlying ability (r=.91) than did the conventional test (r=.89). Both correlations were high and indicated a primarily linear relationship between test scores and ability. Again, there was substantial variability in the test-ability correlations yielded from the 100 samples, indicating caution in the interpretation of results of small-sample (i.e., N=100) simulation studies.

Both the flexilevel and conventional test information functions indicated greatest precision of measurement for "examinees" of near average ability level and decreasing precision with divergence of an examinee's ability from the mean ability level. These findings are in general agreement with those of Lord (1970, 1971d). However, Lord (1971d) also found that the conventional test provided slightly better measurement for ability levels between ±1 standard deviations from the mean, but that the flexilevel test provided better measurement beyond those points, and increased substantially with increasing divergence from the mean. Thus, in Lord's study, the flexilevel test provided more constant precision of measurement across the ability continuum.

In contrast, the results of the present study indicated that the flexilevel test provided more information than did the conventional test at all ability levels between θ=-3.2 and θ=+1.5. Surprisingly, the superiority of the flexilevel test was most apparent for ability levels between θ=-1.0 and θ=0. These results, in combination with the larger standard deviation of flexilevel information values across ability levels, indicate that the flexilevel test provided less constant precision of measurement than did the conventional test. These results are contrary to those of Lord's (1971d) theoretical study of flexilevel testing.

In a simulation study of the two-stage adaptive testing strategy, Betz & Weiss (1974) found that, in agreement with Lord's (1971e) study of two-stage testing, one two-stage test provided relatively constant precision of measurement across the ability continuum; the information function approximated a horizontal line. However, a second two-stage test did not provide constant precision of measurement but rather yielded an information function similar in shape to that of the conventional test although at a higher overall level.

The differences in information values for the conventional and flexilevel tests must also be interpreted in light of differences in the average discriminating power of the test items, since higher item discriminations will generally lead to higher values of information. As was discussed in the section on test construction, the flexilevel test items had a higher mean discrimination ($\bar{a}$=.65) than did the conventional test items ($\bar{a}$=.54), but examinees of relatively low ability would take a more discriminating set of items on this flexilevel test than would examinees of relatively high ability. Thus, where the information provided by the flexilevel and conventional tests was equivalent (about θ=+2.0), the average item discrimination was also equivalent ($\bar{a}$=.54). In the center of the ability distribution, where the flexilevel test showed the greatest advantage over the conventional test, the mean item discrimination for flexilevel was about $\bar{a}$=.69 (again compared to .54 for conventional). At ability levels below θ=-1.5, the flexilevel test still provided more information, but somewhat less than would be expected considering that the mean item discrimination was about $\bar{a}$=.74 or .75.

If item discrimination were the *only* factor influencing information values, the flexilevel test should have the highest values at the lowest ability levels. Thus, while other factors can be assumed to be operating, it does seem that some of the difference between flexilevel and conventional test information values may be attributable to differences in mean item discriminations. Further research using flexilevel tests in which mean item discriminations are equivalent for examinees of all ability levels and are equal to those of conventional tests will be necessary to separate the effects of item discrimination from those of the characteristics of the testing strategies in influencing the overall level and shape of test information functions.

Finally, the flexilevel test provided higher levels of information at lower ability levels than at higher ability levels. While this difference may be due to differences in item discrimination, it contradicts previous findings by Lord regarding the effects of guessing on measurement effectiveness. Lord (1971c) found that guessing had most adverse effects on the measurement effectiveness of both conventional and adaptive tests when examinee ability was low. In a conventional test, low ability examinees receive items which are, for the most part, too difficult for them; thus, their only chance to answer correctly is through guessing. In the flexilevel test, however, fewer items should be too difficult for the low ability examinee, so guessing should be reduced, leading to less measurement error. This hypothesis is supported by the higher information values for the low ability testees, and by the data on proportion correct in the flexilevel test. Again, further research controlling the factor of discrimination will be necessary to determine whether or not flexilevel and other adaptive testing strategies yield scores which contain less error due to guessing, particularly for low ability examinees.

The failure of the results of the simulation study to agree in all respects with those of Lord's (1971d) theoretical study may also be due to the fact that the latter study assumed hypothetical, ideal items, all having the same discriminating power and having difficulties corresponding to exact desired specifications. The present results, however, were obtained using item parameters obtained from a real item pool; the limitations of the pool permitted only approximations to the item characteristics desired for constructing the flexilevel and conventional tests. Further studies using other real or hypothetical but imperfect item pools would be useful in clarifying the advantages and disadvantages of various testing strategies for use in actual applied assessment situations.

Summary

The results of the studies of flexilevel testing showed that a flexilevel test had significantly greater parallel-forms reliability and a significantly higher relationship to underlying ability than did a conventional test. The test-retest stability of the two tests was equivalent for the total group of examinees, but there was some evidence, both from an analysis of the number of items repeated in the flexilevel test and from a comparison of stability and parallel forms reliability coefficients, that memory effects may be more influential in the stability of conventional test scores than in that of flexilevel test scores. The relationship between flexilevel and conventional test scores (r=.89) in the empirical study was as high as the test-retest stability of either test; the relationship shown in the simulation study (r=.82) was less than the parallel-forms reliability of the flexilevel test (r=.84) but greater than that of the conventional test (r=.80). The flexilevel test provided a higher level of information, i.e., greater precision of measurement, than did the conventional test, but it also yielded less constant precision of measurement for examinees of varying ability levels than did the conventional test. However, the interpretation of differences in information values for the two tests was confounded by differences in item discriminating power. Flexilevel test scores better reflected the underlying normal distribution of ability than did conventional test scores, and there was evidence that the flexilevel test was adapting item difficulties to differences in the ability levels of groups and of individuals. The flexilevel test also appeared to reduce guessing. Further research will be necessary to clarify the relative utility of flexilevel and conventional testing strategies in terms of other psychometric and practical criteria.

REFERENCES

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

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.

Betz, N.E. & Weiss, D.J. Simulation studies of two-stage ability testing. Research Report 74-4, Psychometric Methods Program, Department of Psychology, University of Minnesota, 1974.

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.

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

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

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

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

Glass, G.V. & Stanley, J.C. Statistical methods in education and psychology. Englewood Cliffs, N.J.: Prentice-Hall, 1970.

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

Hansen, D.N., Johnson, B.G., Fagan, R.L., Tam, P. & Dick, W. Computer-based adaptive testing models for the Air Force Technical Training Environment. Phase 1: development of a computerized measurement system for Air Force technical training. AFHRL-TR-74-48, Air Force Human Resources Laboratory, Brooks Air Force Base, Texas, 1974.

Jensema, C. An application of latent trait mental test theory to the Washington Pre-College Testing Battery. Unpublished doctoral dissertation, University of Washington, 1972.

Kocher, A.T. An empirical investigation of the stability and accuracy of flexilevel tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, 1974.

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.

Larkin, K.C. & Weiss, D.J. An empirical comparison of two-stage and pyramidal adaptive ability testing. Research Report 75-1, Psychometric Methods Program, Department of Psychology, University of Minnesota, 1975.

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

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

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

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

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

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

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

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

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

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

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.

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

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

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

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

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

Weiss, D.J. Strategies of adaptive ability measurement.. Research Report 74-5, Psychometric Methods Program, Department of Psychology, University of Minnesota, 1974.

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

Table A-1
Item Reference Numbers and Normal Ogive
Item Parameters for the Flexilevel Test

Item Reference Number[a]	Item Serial Number	Difficulty (*b*)	Discrimination (*a*)	Item Reference Number[a]	Item Serial Number	Difficulty (*b*)	Discrimination (*a*)
121	1	-3.11	.70	655	41	.08	.39
131	2	-2.98	.56	386	42	.14	.70
89	3	-2.82	.67	266	43	.16	.86
198	4	-2.81	.74	264	44	.21	.86
82	5	-2.77	.50	340	45	.30	.78
80	6	-2.55	.79	296	46	.34	.91
184	7	-2.54	.67	111	47	.46	.48
31	8	-2.50	.66	213	48	.65	.29
66	9	-2.32	.80	164	49	.62	.41
95	10	-2.20	.50	656	50	.71	.44
262	11	-2.29	.70	294	51	.79	.70
214	12	-2.08	.42	321	52	.79	.63
34	13	-1.93	.74	216	53	.92	.37
83	14	-1.80	.77	299	54	.98	.52
186	15	-1.65	.92	120	55	1.07	.72
88	16	-1.74	.63	147	56	1.15	.38
199	17	-1.42	.92	217	57	1.25	.43
103	18	-1.34	.89	668	58	1.26	.39
173	19	-1.43	.76	652	59	1.33	.60
47	20	-1.31	.87	152	60	1.40	.55
43	21	-1.21	.90	400	61	1.62	.34
87	22	-1.10	.99	359	62	1.54	.58
109	23	-1.06	.89	319	63	1.49	.62
204	24	-1.15	.73	253	64	1.65	.39
85	25	-1.07	.76	383	65	1.82	.36
123	26	-1.00	.67	273	66	1.79	.49
349	27	- .94	.74	379	67	1.94	.64
130	28	- .85	.75	166	68	2.03	.64
128	29	- .75	.82	672	69	1.89	.85
37	30	- .69	.66	297	70	2.31	.40
91	31	- .59	.83	336	71	2.05	.49
270	32	- .52	.86	309	72	2.47	.48
188	33	- .47	.71	245	73	2.32	.38
145	34	- .41	.59	398	74	2.34	.61
209	35	- .40	.64	385	75	2.35	.42
56	36	- .28	.75	298	76	2.62	.43
329	37	- .21	.86	364	77	3.11	.32
272	38	- .13	.98	388	78	2.86	.43
630	39	- .05	1.31	664	79	2.95	.84
258	40	.00	.41				
					Mean	-.01	.65
					S.D.	1.68	.20

[a] Refers to item numbers used in McBride & Weiss (1974) Appendix A.

Table A-2

Item Reference Numbers and Normal Ogive
Item Parameters for the Conventional Test

Item Reference Numbers[a]	Difficulty (b)	Discrimination (a)
58	-.96	.48
221	-.74	.65
307	-.84	.56
393	-.95	.49
211	-.72	.61
224	-.79	.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

[a] Refers to item numbers used in McBride & Weiss (1974) Appendix A.

Table A-3

Information values from Time 1 and Time 2 administrations, and averages, for flexilevel and conventional tests (equal-frequency distribution)

Level of Ability (θ)	Flexilevel			Conventional		
	Time 1	Time 2	Average	Time 1	Time 2	Average
3.2	.11	.01	.05	1.06	.65	.85
3.0	.33	.11	.22	.03	.01	.02
2.5	1.12	1.03	1.08	1.07	1.18	1.12
2.0	3.47	2.33	2.90	3.37	3.22	3.29
1.5	6.56	4.63	5.60	3.86	4.90	4.38
1.0	6.88	5.04	5.96	4.81	4.03	4.42
.5	6.53	5.60	6.07	6.07	4.53	5.30
.1	6.53	6.93	6.73	3.96	4.96	4.46
- .1	6.48	7.61	7.05	4.31	4.45	4.38
- .5	6.02	5.97	5.99	4.88	3.62	4.25
-1.0	7.18	5.77	6.47	3.71	3.34	3.53
-1.5	4.07	4.92	4.50	3.51	2.51	3.01
-2.0	3.89	3.75	3.82	2.33	2.66	2.50
-2.5	2.22	2.66	2.44	1.38	1.25	1.32
-3.0	2.21	2.52	2.36	.28	.51	.39
-3.2	1.38	1.42	1.40	.08	.22	.15
Mean	4.06	3.77	3.92	2.79	2.63	2.71
S.D.	2.56	2.39	2.44	1.92	1.76	1.81

Note. N=100 per ability level

Table A-4

Smoothed values of the information function for flexilevel and conventional tests within intervals of the continuum of underlying ability (normal distribution of ability levels)

Interval of Ability (θ)	N	Flexilevel	Conventional
3.1 to 3.3	4	.18	.19
2.9 to 3.1	14	.41	.27
2.7 to 2.9	30	.74	.41
2.5 to 2.7	52	1.08	.76
2.3 to 2.5	100	1.48	1.47
2.1 to 2.3	168	2.05	2.25
1.9 to 2.1	192	2.69	3.00
1.7 to 1.9	318	3.40	3.50
1.5 to 1.7	452	3.88	3.81
1.3 to 1.5	596	4.31	4.09
1.1 to 1.3	742	4.60	4.28
.9 to 1.1	1042	4.97	4.46
.7 to .9	1088	5.26	4.50
.5 to .7	1334	5.56	4.56
.3 to .5	1496	5.77	4.63
.1 to .3	1442	5.93	4.64
-.1 to .1	1690	5.99	4.60
-.3 to -.1	1548	5.90	4.43
-.5 to -.3	1550	5.72	4.24
-.7 to -.5	1264	5.45	3.99
-.9 to -.7	1156	5.26	3.78
-1.1 to -.9	948	5.05	3.61
-1.3 to -1.1	652	4.87	3.47
-1.5 to -1.3	660	4.55	3.29
-1.7 to -1.5	470	4.18	3.07
-1.9 to -1.7	350	3.89	2.76
-2.1 to -1.9	208	3.74	2.40
-2.3 to -2.1	144	3.66	2.08
-2.5 to -2.3	82	3.34	1.75
-2.7 to -2.5	56	2.92	1.49
-2.9 to -2.7	40	3.07	1.08
-3.1 to -2.9	16	3.24	.72
-3.3 to -3.1	12	2.43	.31
Mean		3.81	2.85
S.D.		1.70	1.53

Note. Values obtained using method of "moving averages" (McNemar, 1969, p. 8).

Table A-5

Information values from Time 1 and Time 2 administrations, and averages, for flexilevel and conventional tests (normal distribution of ability) (N=10,000)

Interval of Ability (θ)	N	Flexilevel (Score 2)			Conventional		
		Time 1	Time 2	Average	Time 1	Time 2	Average
3.1 to 3.3	4	.25	.01	.13	.58	*	.58
2.9 to 3.1	14	.24	.06	.15	.04	.42	.23
2.7 to 2.9	30	1.17	.93	1.05	.12	.01	.06
2.5 to 2.7	52	1.13	.67	.90	.43	.41	.42
2.3 to 2.5	100	1.61	.96	1.28	1.93	1.40	1.66
2.1 to 2.3	168	2.35	1.96	2.16	2.02	1.85	1.93
1.9 to 2.1	192	2.42	2.19	2.30	4.23	2.73	3.48
1.7 to 1.9	318	4.13	3.60	3.87	3.48	3.99	3.73
1.5 to 1.7	452	4.24	3.66	3.95	3.46	3.64	3.55
1.3 to 1.5	596	3.99	4.61	4.30	3.97	4.54	4.26
1.1 to 1.3	742	4.14	4.81	4.47	4.05	4.35	4.20
.9 to 1.1	1042	5.08	5.35	5.21	4.36	5.19	4.77
.7 to .9	1088	5.22	4.88	5.05	4.27	4.45	4.36
.5 to .7	1334	5.74	5.69	5.72	4.54	4.29	4.42
.3 to .5	1496	5.57	6.10	5.84	4.78	4.96	4.87
.1 to .3	1442	5.84	5.96	5.90	4.58	4.53	4.55
-.1 to .1	1690	6.51	5.74	6.13	5.18	4.39	4.78
-.3 to -.1	1548	5.74	6.37	6.01	4.37	4.51	4.44
-.5 to -.3	1550	5.82	5.86	5.84	4.26	4.24	4.25
-.7 to -.5	1264	5.26	5.26	5.26	4.28	3.63	3.96
-.9 to -.7	1156	5.44	5.21	5.33	3.87	3.70	3.78
-1.1 to -.9	948	4.83	4.93	4.88	3.64	3.33	3.48
-1.3 to -1.1	652	4.93	5.46	5.20	3.78	3.33	3.56
-1.5 to -1.3	660	4.52	4.51	4.51	3.85	2.87	3.36
-1.7 to -1.5	470	4.14	4.14	4.14	3.39	2.80	3.10
-1.9 to -1.7	350	3.72	3.92	3.82	2.98	2.50	2.74
-2.1 to -1.9	208	3.06	3.57	3.32	2.97	2.11	2.54
-2.3 to -2.1	144	5.10	3.16	4.13	2.57	1.40	1.98
-2.5 to -2.3	82	4.69	2.61	3.65	1.46	1.50	1.48
-2.7 to -2.5	56	2.50	3.11	2.80	3.10	.94	2.02
-2.9 to -2.7	40	1.89	2.03	1.96	1.26	.33	.80
-3.1 to -2.9	16	1.22	3.80	2.51	.71	.98	.84
-3.3 to -3.1	12	12.52	2.41	7.46	.29	.01	.15
Mean		4.09	3.74	3.92	2.99	2.79	2.86
S.D.		2.34	1.85	1.88	1.56	1.63	1.57

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

DISTRIBUTION LIST

Navy

4 Dr. Marshall J. Farr, Director
Personnel and Training Research Programs
Office of Naval Research (Code 458)
Arlington, VA 22217

1 ONR Branch Office
495 Summer Street
Boston, MA 02210
ATTN: Research Psychologist

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

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

6 Director
Naval Research Laboratory
Code 2627
Washington, DC 20390

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

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

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

1 Chief of Naval Reserve
Code 3055
New Orleans, LA 70146

1 AFHRL/PE
Stop 63
Lackland AFB, Texas 78236

1 Navy Personnel Research and
Development Center
Code 9041
San Diego, California 92152
Attn: Dr. J. D. Fletcher

1 Dr. Leo Miller
Naval Air Systems Command
AIR-413E
Washington, DC 20361

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

1 Chief
Bureau of Medicine & Surgery
Research Division (Code 713)
Washington, DC 20372

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

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

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

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

1 Director, Navy Occupational Task
Analysis Program (NOTAP)
Navy Personnel Program Support
Activity
Building 1304, Bolling AFB
Washington, DC 20336

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

1 Department of the Navy
Office of Civilian Manpower Management
Code 263
Washington, DC 20390

1 Dr. Robert Smith
Chief of Naval Operations (OP-987E)
Department of the Navy
Washington, DC 20350

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

1 Commander, Navy Recruiting Command
4015 Wilson Boulevard
Arlington, VA 22203
ATTN: Code 015

1 Mr. George N. Graine
Naval Ship Systems Command
SHIPS 047C12
Washington, DC 20362

1 Chief of Naval Technical Training
Naval Air Station Memphis (75)
Millington, TN 38054
ATTN: Dr. Norman J. Kerr

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

1 Dr. Alfred F. Smode, Staff Consultant
Training Analysis & Evaluation Group
Naval Training Equipment Center
Code N-OOT
Orlando, FL 32813

1 Dr. Hanns H. Wolff
Technical Director (Code N-2)
Naval Training Equipment Center
Orlando, FL 32813

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

1 Dr. Charles Cory
Navy Personnel R&D Center
San Diego, CA 92152

5 Navy Personnel R&D Center
San Diego, CA 92152
ATTN: Code 10

1 D. M. Gragg, CAPT, MC, USN
Head, Educational Programs Development
Department
Naval Health Sciences Education and
Training Command
Bethesda, MD 20014

Army

1 Headquarters
U.S. Army Administration Center
Personnel Administration Combat
Development Activity
ATCP-HRO
Ft. Benjamin Harrison, IN 46249

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

1 Commandant
United States Army Infantry School
ATTN: ATSH-DET
Fort Benning, GA 31905

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

1 Dr. Frank J. Harris
U.S. Army Research Institute
1300 Wilson Boulevard
Arlington, VA 22209

1 Dr. Stanley L. Cohen
U.S. Army Research Institute
1300 Wilson Boulevard
Arlington, VA 22209

1 Dr. Ralph Dusek
U.S. Army Research Institute
1300 Wilson Boulevard
Arlington, VA 22209

1 Dr. Milton Maier
U.S. Army Research Institute
1300 Wilson Boulevard
Arlington, VA

1 Dr. Ralph Canter
U.S. Army Research Institute
1300 Wilson Boulevard
Arlington, VA 22209

1 Dr. J.E. Uhlaner, Technical Director
U.S. Army Research Institute
1300 Wilson Boulevard
Arlington, VA 22209

1 Dr. Joseph Ward
U.S. Army Research Institute
1300 Wilson Boulevard
Arlington, VA 22209

1 HQ USAREUR & 7th Army
ODCSOPS
USAREUR Director of GED
APO New York 09403

Air Force

1 Research Branch
AF/DPMYAR
Randolph AFB, TX 78148

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

1 AFHRL/DOJN
Stop #63
Lackland AFB, TX 78236

1 Dr. Robert A. Bottenberg (AFHRL/SM)
Stop #63
Lackland AFB, TX 78236

1 Dr. Martin Rockway (AFHRL/TT)
Lowry AFB
Colorado 80230

1 Major P.J. DeLeo
Instructional Technology Branch
AF Human Resources Laboratory
Lowry AFB, CO 80230

1 AFOSR/NL
1400 Wilson Boulevard
Arlington, VA 22209

1 Dr. Sylvia R. Mayer (MCIT)
Headquarters Electronic Systems Division
LG Hanscom Field
Bedford, MA 01730

1 CAPT Jack Thorpe, USAF
Flying Training Division (HRL)
Williams AFB, AZ 85224

Marine Corps

1 Mr. E.A. Dover
Manpower Measurement Unit (Code MPI)
Arlington Annex, Room 2413
Arlington, VA 20380

1 Commandant of the Marine Corps
Headquarters, U.S. Marine Corps
Code MPI-20
Washington, DC 20380

1 Director, Office of Manpower Utilizatio
Headquarters, Marine Corps (Code MPU)
MCB (Building 2009)
Quantico, VA 22134

1 Dr. A.L. Slafkosky
Scientific Advisor (Code RD-1)
Headquarters, U.S. Marine Corps
Washington, DC 20380

1 Chief, Academic Department
Education Center
Marine Corps Development and
Education Command
Marine Corps Base
Quantico, VA 22134

Coast Guard

1 Mr. Joseph J. Cowan, Chief
Psychological Research Branch (G-[illegible]
U.S. Coast Guard Headquarters
Washington, DC 20590

Other DOD

1 Lt. Col. Henry L. Taylor, USAF
Military Assistant for Human Resources
OAD (E&LS) ODDR&E
Pentagon, Room 3D129
Washington, DC 20301

1 Mr. William J. Stormer
DOD Computer Institute
Washington Navy Yard, Building 175
Washington, DC 20374

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

Other Government

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

1 Dr. William Gorham, Director
Personnel Research and Development
Center
U.S. Civil Service Commission
1900 E. Street, N.W.
Washington, DC 20415

1 Dr. Vern Urry
Personnel Research and Development
Center
U.S. Civil Service Commission
1900 E. Street, N.W.
Washington, DC 20415

1 Dr. Eric McWilliams, Program Manager
Technology and Systems, TIE
National Science Foundation
Washington, DC 20550

1 Dr. Andrew R. Molnar
Technological Innovation
National Science Foundation
Washington, DC 20550

1 U.S. Civil Service Commission
Federal Office Bldg.
Chicago Regional Staff Div.
Attn: C. S. Winiewicz
Regional Psychologist
230 So Dearborn St.
Chicago, IL 60604

Miscellaneous

1 Dr. Scarvia B. Anderson
Educational Testing Service
17 Executive Park Drive, N.E.
Atlanta, GA 30329

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

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

1 Dr. Gerald V. Barrett
University of Akron
Department of Psychology
Akron, OH 44325

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

1 Dr. Ronald P. Carver
School of Education
University of Missouri - Kansas City
Kansas City, Missouri 64110

1 Century Research Corporation
4113 Lee Highway
Arlington, VA 22207

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

1 Dr. Allan M. Collins
Bolt Beranek and Newman, Inc.
50 Moulton Street
Cambridge, MA 02138

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

1 Dr. Norman R. Dixon
Room 170
190 Lothrop Street
Pittsburgh, PA 15260

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

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

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

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

1 Dr. Edwin A. Fleishman
American Institutes for Research
Foxhall Square
3301 New Mexico Avenue, N.W.
Washington, DC 20016

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

1 Mr. Harry H. Harman
Educational Testing Service
Princeton, NJ 08540

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

1 Dr. M.D. Havron
Human Sciences Research, Inc.
7710 Old Spring House Road
West Gate Industrial Park
McLean, VA 22101

1 HumRRO
Division No. 3
P.O. Box 5787
Presidio of Monterey, CA 93940

1 HumRRO
Division No. 4, Infantry
P.O. Box 2086
Fort Benning, GA 31905

1 HumRRO
Division No. 5, Air Defense
P.O. Box 6057
Fort Bliss, TX

1 HumRRO
Division No. 6, Library
P.O. Box 428
Fort Rucker, AL 36360

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

1 Dr. Milton S. Katz
MITRE Corporation
Westgate Research Center
McLean, VA 22101

1 Dr. Steven W. Keele
University of Oregon
Department of Psychology
Eugene, OR 97403

1 Dr. David Klahr
Carnegie-Mellon University
Department of Psychology
Pittsburgh, PA 15213

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

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

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

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

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

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

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

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

1 Dr. C. Harold Stone
1428 Virginia Avenue
Glendale, CA 91202

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

1 Dr. Benton J. Underwood
Northwestern University
Department of Psychology
Evanston, IL 60201

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

1 Mr. Edmond Marks
405 Old Main
Pennsylvania State University
University Park, PA 16802

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

1 Dr. Donald A. Norman
University of California, San Diego
Center for Human Information Processing
LaJolla, CA 92037

Previous Reports in this Series

73-1. Weiss, D.J. & Betz, N.E. Ability Measurement: Conventional or Adaptive? February 1973 (AD 757788).

73-2. Bejar, I.I. & Weiss, D.J. Comparison of Four Empirical Differential Item Scoring Procedures. August 1973.

73-3. Weiss, D.J. The Stratified Adaptive Computerized Ability Test. September 1973 (AD 768376).

73-4. Betz, N.E. & Weiss, D.J. An Empirical Study of Computer-Administered Two-stage Ability Testing. October 1973 (AD 768993).

74-1. DeWitt L.J. & Weiss, D.J. A Computer Software System for Adaptive Ability Measurement. January 1974 (AD 773961).

74-2. McBride, J.R. & Weiss, D.J. A Word Knowledge Item Pool for Adaptive Ability Measurement. June 1974 (AD 781894).

74-3. Larkin, K.C. & Weiss, D.J. An Empirical Investigation of Computer-Administered Pyramidal Ability Testing. July 1974 (AD 783553).

74-4. Betz, N.E. & Weiss, D.J. Simulation Studies of Two-stage Ability Testing. October 1974 (AD A001230)

74-5. Weiss, D.J. Strategies of Adaptive Ability Measurement. December 1974. (AD A004270)

75-1. Larkin, K.C. & Weiss, D.J. An Empirical Comparison of Two-stage and Pyramidal Adaptive Ability Testing. February 1975. (AD A006733)

75-2. McBride, J.R. & Weiss, D.J. TETREST: A FORTRAN IV program for calculating tetrachoric correlations. March 1975. (AD A007572)

AD Numbers are those assigned by the Defense Documentation Center, for retrieval through the National Technical Information Service

Copies of these reports are available, while supplies last, from:

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