AD-A076 299 MISSOURI UNIV-COLUMBIA TAILORED TESTING RESEARCH LAB F/G 5/10
PROBLEMS IN APPLICATION OF LATENT TRAIT MODELS TO TAILORED TEST--ETC(U)
SEP 79 W R KOCH , M D RECKASE N00014-77-C-0097
UNCLASSIFIED RR-79-1 NL

| OF |

AD
A076299

END
DATE
FILMED
11-79
DDC

Problems in Application of Latent Trait Models to Tailored Testing

William R. Koch
and
Mark D. Reckase

Research Report 79-1
September 1979

Tailored Testing Research Laboratory
Educational Psychology Department
University of Missouri
Columbia, MO 65211

Prepared under contract No. N00014-77-C-0097, NR150-395
with the Personnel and Training Research Programs
Psychological Sciences Division
Office of Naval Research

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

1

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
Research Report 79-1

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
6 Problems in Application of Latent Trait Models to Tailored Testing

5. TYPE OF REPORT & PERIOD COVERED
9 Technical Rep[redacted]t.,

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
10 William R. Koch [redacted] Mark D. Reckase

8. CONTRACT OR GRANT NUMBER(s)
15 N00014-77-C-0097

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Department of Educational Psychology
University of Missouri-Columbia
Columbia, Missouri 65201

410 502

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11 27 Sep 79

11. CONTROLLING OFFICE NAME AND ADDRESS
Personnel and Training Research Programs
Office of Naval Research
Arlington, Virginia 22217

12. REPORT DATE
September 27, 1979

13. NUMBER OF PAGES
53

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
12 54 14 RR-79-1

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approval 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
Achievement Testing
Latent Trait Models
Rasch Model
Tailored Testing
Computerized Testing
Adaptive Testing

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

Computerized tailored testing procedures have been successfully applied in the past to the measurement of aptitude or ability. The latent trait models employed in these procedures make the basic assumption that the underlying latent trait being measured is unidimensional. However, achievement tests are commonly found to measure several factors. The purpose of the present research was to study the effects of using tailored tests for achievement measurement, knowing that the unidimensionality assumption

→ next page

420 502

#20(Cont.)

cont. would be violated. Of equal importance to the study was a comparison of the one- and three-parameter logistic models to each other as well as to a traditional paper-and-pencil achievement test. A total of 110 undergraduate students enrolled in an introductory educational psychology and measurement course at the University of Missouri-Columbia served as examinees for the study. A counterbalanced test-retest design was employed in which there were two separate test sessions one week apart for each examinee, with both the one- and three-parameter tests administered at each session. The tailored tests were administered on Applied Digital Data Systems Consul 980 cathode ray tube terminals which were connected to an IBM 370/168 computer through a timesharing system. Relative efficiency curves, test-retest reliability coefficients, goodness of fit of the models, descriptive statistics, content validity, and the correlation of the tailored test ability estimates with the traditional course exam scores were used to compare the models. Item pools were constructed through the use of linking procedures to place item parameters from different test calibrations onto the same scale. During the tailored test, items were selected for administration based on the information function, and maximum likelihood ability estimation was employed. In addition, an attitude survey was administered after each test session to determine student attitudes toward the tailored tests. The results of the study indicated that neither tailored test procedure performed as well as the traditional course exam in terms of reliability. However, the three-parameter procedure had higher test information and better fit of observed responses to the model than the one-parameter procedure. Neither the one-parameter nor the three-parameter tailored tests yielded satisfactory content validity. The attitude scale results indicated generally favorable student attitudes toward tailored testing.

CONTENTS

Problems in Application of Latent Trait

Models to Tailored Testing

Tailored testing has frequently been proposed as an innovative solution to many age-old measurement problems. In particular, tailored testing procedures can theoretically alleviate many commonly encountered problems with conventional, paper-and-pencil multiple choice tests. One problem with conventional tests, in which all the examinees are administered the same questions, is that test items are often of inappropriate difficulty for many examinees. An examinee with low ability may be frustrated by the difficult items on the test, and therefore, will resort to random guessing or to item omissions. On the other hand, an examinee with a high ability level will often find many test items to be too easy and unchallenging. In general, there is a tendency for conventional tests to be most appropriate and accurate for measuring the average examinee. This tendency is reflected by the fact that the standard error of measurement of a test is usually higher at the extremes than in the middle of the ability range. The result of imprecise measurement, of course, is lower overall test reliability. Other commonly cited problems with conventional tests include time limit pressures and effects of test administration differences (Weiss, 1974).

Tailored testing procedures (Lord, 1970) have been developed in an attempt to alleviate these and other problems with conventional tests, but tailored testing strategies may often be accompanied by a whole new host of problems. The purpose of this report is to describe some of the difficulties which became evident while conducting tailored testing research at the University of Missouri-Columbia. First, however, it may be helpful to briefly discuss the rationale behind tailored testing and some of its primary characteristics.

The basic goal of tailored testing is to minimize the errors of measurement when estimating an examinee's ability or achievement level. As such, a primary distinguishing feature of tailored testing is its attempt to administer test items of appropriate difficulty level to each examinee. That is, rather than administering the same set of test items to all examinees, the tailored testing procedures attempt to "tailor make" the test for each individual. This is accomplished by selecting items for administration that maximize the information about an examinee's estimated ability level, resulting in efficient measurement that facilitates the control of test errors.

Tailored testing is often based on item characteristic curve (ICC) theory (Lord, 1952; Lord and Novick, 1968) which involves relatively sophisticated mathematical models. In order to implement tailored testing it is usually necessary to utilize computer capabilities for several steps. First, tailored testing requires a precalibrated pool of items for the selection of test items to be administered. Calibration of items is usually accomplished by submitting item response data from some conventional test to one of several existing latent trait calibration programs (Wright and Panchapakesan, 1969; Wood, Wingersky, and Lord, 1976; and Urry, 1975) in order to obtain item parameter estimates such as difficulty, discrimination, and guessing indexes.

Another required step is the development of a computer program to operate the tailored testing procedure on an interactive basis with the examinee. In developing this program, many decisions must be made as to the operational characteristics of the test itself: (a) the entry point into the item pool (the first item administered), (b) the ability estimation procedure to be utilized (usually either a Bayesian or maximum likelihood technique), (c) the method used to select successive items, given responses on the previous items, and (d) a stopping rule to terminate the test.

As might be expected, numerous problems may arise that must be dealt with in order to establish tailored testing as a viable alternative to conventional testing. In particular, the item calibration and ability estimation phases of tailored testing present special difficulties. These will be considered in greater detail later in this report, but it will suffice for now to note that, first, sample size is an important determinant of item calibration quality (Reckase, 1977). Moreover, calibration weaknesses may be compounded when data from several small sample calibrations are linked together using items in common to form a larger item pool. Another problem that may occur under certain circumstances is the nonconvergence of ability estimation procedures. Finally, some of the assumptions of the latent trait models may be violated in tailored testing procedures, resulting in problems when, for example, an extension is made from ability testing to applications in achievement testing.

Latent Trait Models

The Rasch (1960), or one-parameter logistic (1PL) model, has been thoroughly described by Wright (1977). In general, the 1PL model requires only one ability parameter, θ_j, for each person and one item difficulty parameter, b_i, for each item in order to represent the interaction between an examinee and a test item. The exponential form of the 1PL model is

$$P(u_{ij}) = \frac{\exp(u_{ij}(\theta_j - b_i))}{1 + \exp(\theta_j - b_i)} \qquad (1)$$

where u_{ij} is the score (0 or 1) on Item i by Person j, θ_j and b_i are as defined above, and $P(u_{ij})$ is the probability that u_{ij} is equal to 0 or 1.

In contrast, the three-parameter logistic (3PL) model presented by Birnbaum (1968) requires the estimation of three item parameters to represent the interaction between test items and examinees. The model is given by

$$P_{ij} = P(u_{ij} = 1) = c_i + (1 - c_i) \frac{\exp(Da_i(\theta_j - b_i))}{1 + \exp(Da_i(\theta_j - b_i))} \quad (2)$$

where $P(u_{ij} = 1)$ is the probability of a correct response by Person j to Item i; c_i is the guessing parameter for Item i; D is a scaling constant equal to 1.7; a_i is the item discrimination parameter; b_i is the item difficulty parameter; and θ_j is the ability parameter for Person j. The probability of an incorrect response, Q_{ij} is defined simply as $1 - P_{ij}$.

Both models have in common the assumptions that the items are scored dichotomously, that the latent trait being measured by the items is unidimensional, that the model describes the interaction between a person and an item, and that local independence holds (Lord and Novick, 1968). This last assumption simply means that the probability of a response (correct or incorrect) to any given item on a test is unaffected by any previous response.

The unidimensionality assumption has particular relevance when considering tailored testing applications to ability tests compared to achievement tests. In the former case, factor analytic procedures usually yield one dominant factor being measured by the test items. Certainly this is the case for ability measures such as verbal or quantitative aptitude, and often is the case for intelligence tests.

On the other hand, achievement tests are usually constructed with multidimensional measurement as a primary goal. Since most achievement tests are based on the objective of sampling distinct content areas or domains, multidimensionality inevitably seems to be built into the tests. With this being the case, the unidimensional assumption of latent trait measurement needs to be examined for achievement test applications of tailored testing. The present study brings evidence to bear on this issue and will be discussed in detail later. However, it is convenient as a basis for comparison to first summarize the results of a previous study reported on tailored testing applied to unidimensional vocabulary ability measurement (Koch and Reckase, 1978).

Vocabulary Tailored Testing Study

The purpose of the study was to compare the 1PL and 3PL models in a tailored testing application to vocabulary ability measurement. A counterbalanced test-retest design was employed in which there were two separate test sessions one week apart for each examinee, with both the 1PL and 3PL tests administered at each session. The calibration programs used to obtain item parameter estimates for the 72 items contained in the vocabulary pool were the Wright and Panchapakesan (1968) program for the 1PL model and the LOGIST program (Wood, Wingersky, and Lord, 1976) for the 3PL model. Test items were selected for administration to maximize the information function (Birnbaum, 1968) for the maximum likelihood ability estimates.

In general the results demonstrated that tailored tests based on either of these two latent trait models could be successfully applied to vocabulary ability measurement. However, there were several specific areas where one tailored testing method performed better than the other. For example, the 3PL test was found not only to have more total test information than the 1PL test, but also to have a better fit between the empirically obtained responses and those predicted by the model.

In regard to reliability, the 3PL procedures resulted in a significantly higher reliability coefficient than the 1PL test. The values, which reflected a combination of test-retest and equivalent forms reliability, were $r = .77$ and $r = .61$, respectively. However, it must be emphasized that the 3PL procedure, in conjunction with maximum likelihood ability estimation, failed to converge to ability estimates in nearly one-third of the tailored tests. With these nonconvergence cases included in the reliability calculation, the correlation coefficient for the 3PL tests dropped to $r = .36$. With maximum likelihood scoring being a major technique for ability estimation, the nonconvergence phenomenon constituted a serious problem. The hypothesis was forwarded that the nonconvergence was due to the item pool being too difficult overall for numerous examinees. It is important to note that nonconvergence of ability estimation never occurred in conjunction with the 1PL model.

Tailored Achievement Testing

With the exception of the research reported by the Psychometric Methods Program at the University of Minnesota, virtually nothing has been published in the literature in regard to the application of tailored testing to achievement measurement. Although frequently treated as if they are highly similar in approach (e.g. Bejar, Weiss and Kingsbury, 1977), the tailored measurement of ability and achievement can present quite different problems. Ability tests commonly make use of high quality, unidimensional pools of items. In contrast, since achievement tests are constructed to measure distinct content areas, the item pools are usually multidimensional.

However, the biology achievement test studied by Bejar, et al. (1977) was found to have only one dominant factor. With this being the case, it was not surprising that the calibration of the item pool with the unidimensional item characteristic curve (ICC) model proved to be adequate. In fact, the use of the ICC model with the biology achievement test would not be expected to be much different than unidimensional aptitude or ability tests.

A more interesting application of ICC theory was reported by Brown and Weiss (1977) in which an adaptive testing procedure was used for an achievement test which had several content areas. This research nicely demonstrated that an adaptive testing strategy utilizing inter-subtest branching substantially reduced the total test length while, at the same time, providing equal precision of measurement compared with the conventional achievement test battery. In addition, the correlations between the adaptive subtest scores and the conventional subtest scores were quite high, with the adaptive subtest information curves being nearly identical to the conventional subtest information curves.

However, even this application of adaptive testing to multidimensional achievement measurement (Brown and Weiss, 1977) did not address the issue of the robustness of ICC theory with respect to the violation of the unidimensionality assumption. This was due to the fact that each subtest or content area was calibrated separately, rather than having one calibration of a multidimensional item pool. Nor was there any attempt to examine another crucial aspect of achievement testing, namely content validity. This is the question of whether or not the tailored test samples the various content areas in the same proportions or to the same degree as the conventional test, assuming that the conventional test was also constructed to have adequate content validity. The current study provided an opportunity to investigate both the robustness of the ICC model and the content validity of tailored testing when applied to achievement measurement.

Method

Item Pool Construction

Calibration The items calibrated for use in the study were obtained from a series of classroom achievement tests which were administered as part of an undergraduate course in educational measurement. Response data were collected from a total of 11 separate 50 item multiple choice exams, most having 4 alternatives per item, covering the content area of educational evaluation techniques. All of the tests were calibrated with both the Wright and Panchapakesan (1969) program and the LOGIST program (Wood, Wingersky, and Lord, 1976) which yielded the 1PL and 3PL item parameter estimates, respectively. The sample sizes ranged from 96 examinees to 314 examinees, although most of the tests had sample sizes of about 200 (see Appendix A-1).

The classroom tests themselves had been produced according to traditional achievement test construction principles. Items were included on the exams if they had moderate to high point biserial discrimination indices, and in such a manner that the average test difficulties were close to .75. Being achievement tests, a table of specifications was used to construct the tests to match course objectives. This meant that separate content areas were identified, behavioral objectives were written at several taxonomic levels for each area, and weights were assigned such that the relative emphasis for different course topics was reflected in the achievement measurement. (See Appendix A-2 for a detailed table of specifications.) KR-20 reliabilities for the exams were found to be consistently in the range of from +.60 to +.80.

Linking Since all of the achievement tests had numerous items in common across tests, item calibration linkings were performed in order to form a large item pool for tailored testing. In this procedure the goal was to link all the separate item parameter calibrations into one final set of item parameters such that parameter estimates obtained from different samples were put onto a single scale. Of course it would be more convenient to have a single large sample of examinees (say 1,000 or more) to which a single test of 150 or more items could be administered. In this latter situation, the need for item parameter linking would be eliminated, and more stable item parameter estimates would be obtained.

Unfortunately, in the typical classroom situation it is rare to have more than 100 examinees taking a single test at one point in time. Moreover, for test security reasons, it is usually necessary to construct a new form of the exam for each new class, although numerous items may overlap. Thus we are confronted with a situation in which many different small sample size calibrations are required to obtain item parameter estimates. One resulting problem is that the parameter scales for each separate calibration are indeterminate.

In the one-parameter program, the zero point on the difficulty scale is arbitrarily set to be at the average item difficulty level for a particular test. With the three-parameter program, the zero point is determined by the average of the ability estimates yielded by the calibration, which is then translated into a zero point on the item difficulty scale. It is easy to see, then, why item parameter estimates turn out to be sample specific. In this regard, it is important to note that these estimates are equivalent within a linear transformation. This means that we still maintain the very desirable attribute of latent trait or ICC models referred to as invariance of item parameters (Lord and Novick, 1968). If the model assumptions are met, then the item parameters will be invariant across different sample calibrations, but only within a transformation of scales. Hence, some form of linking procedure is necessary in order to build a single large calibrated item pool from several test administrations, so that the parameters will be on a common scale.

In order to perform linking of the "b" values (item difficulty parameter estimates) for the one-parameter ICC model, the procedure used in the current study was to identify the items in common among two or more tests and then calculate a mean difficulty value for the common items, separately for each test. One test was then arbitrarily designated as the "calibration base" for the linking. The difference between the mean difficulty for the calibration base and the other test mean difficulty became the scaling constant for linking. This constant was added to all the item difficulty parameters in the second test in order to put them on the same scale as the calibration base item difficulty parameters. For the common items, the transformed parameter estimates were then combined with the base test parameter estimates using a weighted average procedure. Essentially, this procedure amounted to what has been called a major axis scaling procedure (Reckase, 1979).

Linking procedures for the three-parameter ICC model were somewhat more complicated. The procedure used to link the "a" values (item discrimination parameter estimates) was similar to that used for the one-parameter difficulty values, except that a multiplicative constant was used for the scaling. For the items in common between the two tests, this constant was equal to the ratio of the mean of the "a" values for the calibration base items to the mean of the "a" values for the items on the other test. Multiplying all of the "a" values from the second test by this constant was used to transform them onto the same scale as the "a" values from the calibration base. Again, weighting was used to reflect sample size differences for the two tests.

The linking of the three-parameter model "b" values was accomplished through a linear regression procedure. The "b" values from the test to be linked were regressed on the "b" values from the calibration base for the items in common. The resulting regression equation was then used to obtain new estimates of the "b" parameters for the linked test. For the common items, these new parameter estimates were combined with the base test parameter estimates using a weighted average procedure. Since the "c" values (item guessing parameter estimates) from different tests were already on the same 0 to 1 scale, a simple weighted average technique was used to accomplish the linking.

Table 1 presents the means, standard deviations, and ranges of the item parameter estimates resulting from the calibration and linking procedures described above. In addition, Figures 1-A, 1-B, 1-C, and 1-D present histograms of the distributions of the item parameter estimates in the final 180 item pools for the tailored achievement tests. The item pool used for the one-parameter tailored tests contained the same 180 items as the item pool for the three-parameter tailored tests. The correlation between their respective "b" values was .91.

Table 1

Descriptive Statistics of Item Parameter Estimates for Tailored Testing Item Pools

	One-Parameter Calibration	Three-Parameter Calibration		
	b_i	a_i	b_i	c_i
Mean	.518	.758	-1.764	.238
S. D.	1.505	.720	3.800	.115
Low Value	-3.165	.010	-9.999[a]	.000
High Value	5.437	3.537	21.518	.500
No. of Items	180	180	180	180

[a] This value was an arbitrary lower limit on the 3PL difficulty parameters.

As can be seen in Figures 1-B and 1-D, the distributions of item difficulty values in the item pools were markedly peaked rather than taking on a uniform distribution which would have been preferred. Even more disturbing is the distribution of item discrimination parameters shown in Figure 1-A since nearly two-thirds of the items had "a" values below

Item Parameter Distributions

Figure 1-A

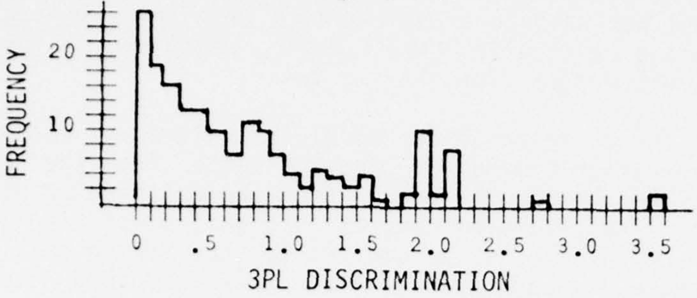
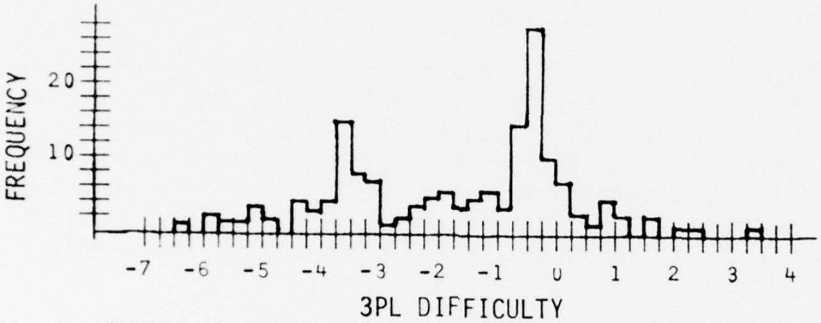


Figure 1-B



Note: *Extreme values of difficulty deleted.*

Figure 1-C

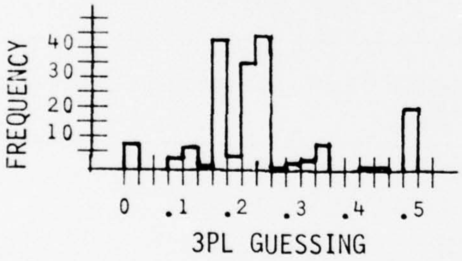
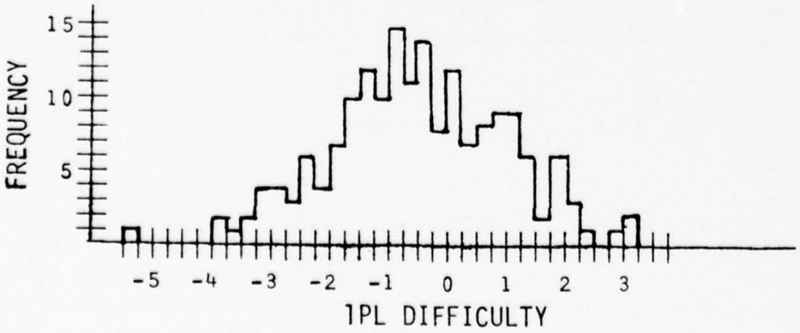


Figure 1-D



.80. Finally, as can be seen in Figure 1-C, approximately thirty items had guessing parameter values above .30, with twenty of these having "c" values close to .50. Clearly the item pool was not an optimal one from a theoretical viewpoint. In practice, it meant that the size of the item pool was severely restricted during the use of the three-parameter tailored testing procedure, as will be seen later. In fact, the 180 item pool was functionally reduced to only 28 items, since items were selected for administration on the basis of their information values.

Tailored Testing Procedures

The three required components of the tailored testing procedure were (a) an item selection routine, (b) an ability estimation technique, and (c) a stopping rule to terminate the test. These components have been described elsewhere (Koch and Reckase, 1978; Patience, 1977), but they will be summarized here.

For both the 1PL and 3PL procedures, items were selected for administration to maximize the value of the information function (Birnbaum, 1968). The information function described the potential contribution of each item to the estimation of a given examinee's ability level. Item information for the 1PL procedure was computed as

$$I(\theta_j) = \frac{\exp[-(\theta_j - b_i)]}{\{1 + \exp[-(\theta_j - b_i)]\}^2} = \psi(\theta_j - b_i) \qquad (3)$$

where $I(\theta_j)$ is the information of Item i at ability level θ for Person j, θ_j and b_i having the same meanings as given in formula 1, and $\psi(x)$ is the logistic probability density function.

For the 3PL procedure, item information was calculated as

$$I(\theta_j) = D^2 a_i^2 \psi[DL_i(\theta_j)] - D^2 a_i P_{ij}(\theta_j)\psi[DL_i(\theta_j) - \log c_i] \qquad (4)$$

where $I(\theta_j)$ is the information as defined above; $L_i(\theta_j) = a_i(\theta_j - b_i)$; $P_{ij}(\theta_j)$ is the probability of a correct response to Item i given ability level θ_j; $\psi(x)$ is the logistic probability density function; and the other parameters have their definitions given previously. The total test information was then simply the sum of the item information (Birnbaum, 1968) given by:

$$I(\theta) = \sum_{i=1}^{n} I(\theta_j) \qquad (5)$$

In the tailored testing procedure, the examinee's initial ability estimate was randomly assigned to be either +.50 or -.50. The first item to be administered was selected such that the information function was maximal at the initial ability estimate. If the examinee answered the first item correctly, the ability estimate was increased by a fixed stepsize (.693) (i.e. a more difficult item). An incorrect response resulted in an ability estimate that was decreased by the stepsize. A fixed stepsize was only used until a maximum likelihood ability estimate could be obtained. In both cases, the item administered was the one with maximum information for the given ability at that point in the test. When at least one correct and one incorrect response were obtained, the ability level of the examinee could then be estimated using an empirical maximum likelihood procedure, with the mode of the likelihood function becoming the new ability estimate. The next item administered was the one in the item pool with maximum information for that ability estimate, with the restriction that no item could be administered more than once during the test.

The tailored tests for both the 1PL and the 3PL procedure cycled through this process until one of two stopping rules was reached: either no item remained in the item pool with an information value for the ability estimate greater than .59 for the 1PL test and .70 for the 3PL test, or a maximum of 20 items had been administered. The information values were different because they were on different theta scales.

Design

The study employed a counterbalanced design in which there were two separate test sessions one week apart for each examinee, with both the 1PL and the 3PL tests administered at each session. The counterbalancing resulted from the reversal of the presentation order of the test models used from one test session to the next. The test-retest feature of the design was planned to facilitate reliability comparisons between the two tailored testing procedures. The tests were arranged so that the examinee could not perceive receiving two tests during each session. The computer program began administering the second test immediately after arriving at an ability estimate from the first test, so there was no pause between them. However, since both item pools were identical in content, the examinees were told that occasionally they would receive the same test item to answer twice. The administration of the tests was accomplished on Applied Digital Data systems (ADDS) Consul 980 cathode ray tube terminals which were connected to an IBM 370/168 through a timesharing system.

Sample

The subjects participating in the study were junior and senior undergraduate students enrolled in an introductory course in measurement and evaluation. Shortly after the students had taken their first course exam, they were asked to volunteer to take other tests over the same material, but in shortened form on a computer terminal. In order to provide some motivation, the instructor informed each student that the tailored tests would be used to assign a course grade if his or her performance was better than the score on the conventional course exam. In addition, all students who participated received extra credit points toward their course grades. A total of 110 students took part.

Attitude Survey

At the end of each tailored test session, the students were requested to respond to a 20 item attitude questionnaire. The scale was identical to the one employed in an earlier vocabulary tailored testing study (Koch and Reckase, 1978) except that a few of the statements were modified slightly to refer to the course achievement test instead of a vocabulary test. All of the statements were written in Likert scale fashion with a five position scale of response alternatives after each item. The items were scored on a scale from 1 to 5 with 1 signifying that the response reflected an unfavorable attitude toward tailored testing and 5 indicating a favorable attitude.

Analyses

The primary research issues in the present achievement test study included comparisons of (a) the respective test-retest reliability coefficients for the 1PL and 3PL tailored testing procedures, (b) the goodness of fit of the two models using mean squared deviations of observed from predicted response data, and (c) the total test information functions for the two tailored testing methods. Also of interest were comparisons of the ability estimates yielded by the two procedures, the content validity of the tailored tests, the correlation of the ability estimates with the conventional course exam, and the attitudes of the students toward tailored testing.

The reliability comparison was based on correlations between the ability estimates yielded by the 1PL and 3PL procedures in the two test sessions. These coefficients were not strictly test-retest reliabilities since no examinee could possibly receive exactly the same tailored test twice, due to different starting points in the item pool and different paths through the pool. However, numerous items were repeated over sessions as a function of the consistency in ability estimation for a person since items were selected from the same pool on the basis of their information values. For the three-parameter procedure, in particular, this meant that highly discriminating items tended to be repeated on the tailored tests. Therefore, the reliability coefficients reflected a mix between test-retest and equivalent forms reliability. The respective reliabilities for the two procedures were compared statistically using a *t*-test based on Fisher's *r* to *z* transformation.

The measure used to determine the goodness of fit of the observed data to the models was the mean squared deviation (MSD) statistic, which was calculated by summing the squared differences for each person between the actual response to an item and the probability of a correct response predicted by the model. These squared differences were computed using the formula

$$MSD_j = \frac{\sum_{i=1}^{n} (u_{ij} - P_{ij})^2}{n} \qquad (6)$$

where MSD_j was the mean squared deviation for Person j, u_{ij} was the actual response to Item i by Person j, P_{ij} was the probability of a correct response to Item i by Person j, and n was the number of items in the tailored test for Person j. A systematic sample of 29 examinees was analyzed to compare the 1PL and 3PL tests using the MSD statistic as the dependent variable in a t-test. The sampling was systematic rather than random to insure that the fit comparison covered the whole range of ability estimates yielded by the tailored tests.

The total test information analyses were performed to compare the 1PL and 3PL procedures in terms of relative efficiency (Birnbaum, 1968). The relative efficiency was the ratio of information provided by each procedure's tailored test (see equations 3, 4, and 5) to the information provided by the traditional 50 item paper-and-pencil course exam. A plot was drawn of the information functions of the tailored versus the traditional test to facilitate comparisons. Again, the plot was constructed based on a selected sample of cases across the whole range of tailored testing ability estimates.

Other data analyses included a series of correlations among several variables that were incorporated in the study. For example, the correlations of the ability estimates yielded by the 1PL and 3PL tailored testing sessions were intercorrelated. Also, the correlation was calculated between tailored testing ability estimates and other criteria of performance. As suggested by Lord (1979) estimated true scores were used for the correlations. The criteria included the students' scores on the traditional course exam over the same content, as well as course exam scores on other content areas. The purposes of these correlations were twofold: to determine the degree to which the two tailored testing procedures were measuring the same thing and to see if one procedure correlated better than the other with the outside criteria. Descriptive statistics from the two tailored testing procedures were also compiled, such as average test length, average test difficulty, number of items actually utilized from the item pools, etc.

Since a major concern of classroom or course achievement tests is content validity, a series of analyses was conducted to determine the degree to which both the item pools and the tailored tests accurately represented the proper measurement of the course objectives. In constructing the course exams, a table of specifications was used to insure proportionate weighting of test items to specified content areas of the course material. The question was whether or not the item pools and the tailored tests themselves also remained faithful to the desired weighting of content areas. A set of chi square analyses were run to measure the goodness of fit between desired and observed item sampling in this respect.

Another question investigated in the present study was the factor structure of the traditional course exam. Since we have argued that achievement tests routinely measure several dimensions, the response data to the conventional course exam was submitted to a principal components analysis with varimax rotation to determine its structure.

A final set of analyses was conducted on the attitude data collected from the questionnaire that was administered after each test session.

Several types of factor analyses were run on the response data such as principal components, image covariance, and alpha factor analysis, using varimax, oblimax, and binormamin rotations. Once the factor structure was determined, attempts were made to label the factors and to compare them with the factors that emerged from the previous administration of the scale (Koch and Reckase, 1978). Coefficient alpha reliabilities were calculated for each factor, as well as for the total scale.

Frequencies of responses to the five scale positions for each item were tabulated for both test sessions. The purpose was to provide a summary table of student attitudes toward the tailored testing procedures. Also, a multivariate analysis of variance was performed to measure any changes in attitudes from one test session to the next.

Results

Goodness of Fit

In Table 2 are presented the results for the MSD statistic used in the goodness of fit comparison of the 1PL and 3PL models. The computed MSD values for 29 cases for each model are shown, along with the means, standard deviations, and the results of a dependent t-test analysis of the data. The results indicated that the MSD statistic was significantly smaller for the 3PL tailored testing procedure ($p<.01$), reflecting better fit of the 3PL model to the observed responses.

Information Function Analyses

The relative efficiency comparison of the total test information for the 1PL and 3PL procedures is shown in Figure 2. The horizontal broken line indicates the information of the traditional 50 item achievement test which was administered in class as the standard for comparing these two types of tailored tests. However, the ability scale used for plotting the 1PL relative efficiency curve is not the same as that for the 3PL relative efficiency curve. (In general, the ability scale for the 1PL model will not be the same as that for the 3PL model.) Even so, a subjective visual comparison of the plots is possible.

In general, the plots indicate that neither tailored test procedure was as informative as the conventional course exam. However, the relative information of the 3PL procedure came substantially closer to the traditional paper-and-pencil exam than did the 1PL tailored tests. This finding was in contrast to the vocabulary tailored testing study results (Koch and Reckase, 1978) which showed the 3PL procedure to have more information than the conventional test, while the 1PL procedure had almost as much information as the conventional test. The overall shape of the information relative efficiency curve was somewhat irregular for the 1PL tests, but it was peaked for the 3PL tests. Also, the 1PL procedure had its highest relative efficiency at the upper extremes of ability where very few examinees were classified, while the 3PL tests were most

Table 2

Goodness of Fit Comparison
Using the MSD Statistic

Observations	One-Parameter MSD	Three-Parameter MSD
1	.2136	.1115
2	.2156	.2745
3	.2015	.1507
4	.2063	.1808
5	.2119	.1471
6	.1902	.1216
7	.1917	.0979
8	.2184	.2207
9	.2207	.2047
10	.2051	.2311
11	.1677	.1642
12	.1990	.2086
13	.1991	.1897
14	.2099	.2132
15	.1775	.1515
16	.2064	.0943
17	.2216	.0966
18	.1797	.1166
19	.2094	.1723
20	.2198	.2554
21	.1560	.0962
22	.2133	.1210
23	.2040	.1012
24	.2182	.2841
25	.2034	.0762
26	.2434	.2061
27	.1962	.0672
28	.2175	.1620
29	.2168	.2649
$\overline{X}$	.2046	.1649
$S_{\overline{X}}$	.0426	.0701

$\underline{t}_{(28)} = 3.727$ $(\underline{p} < .01)$

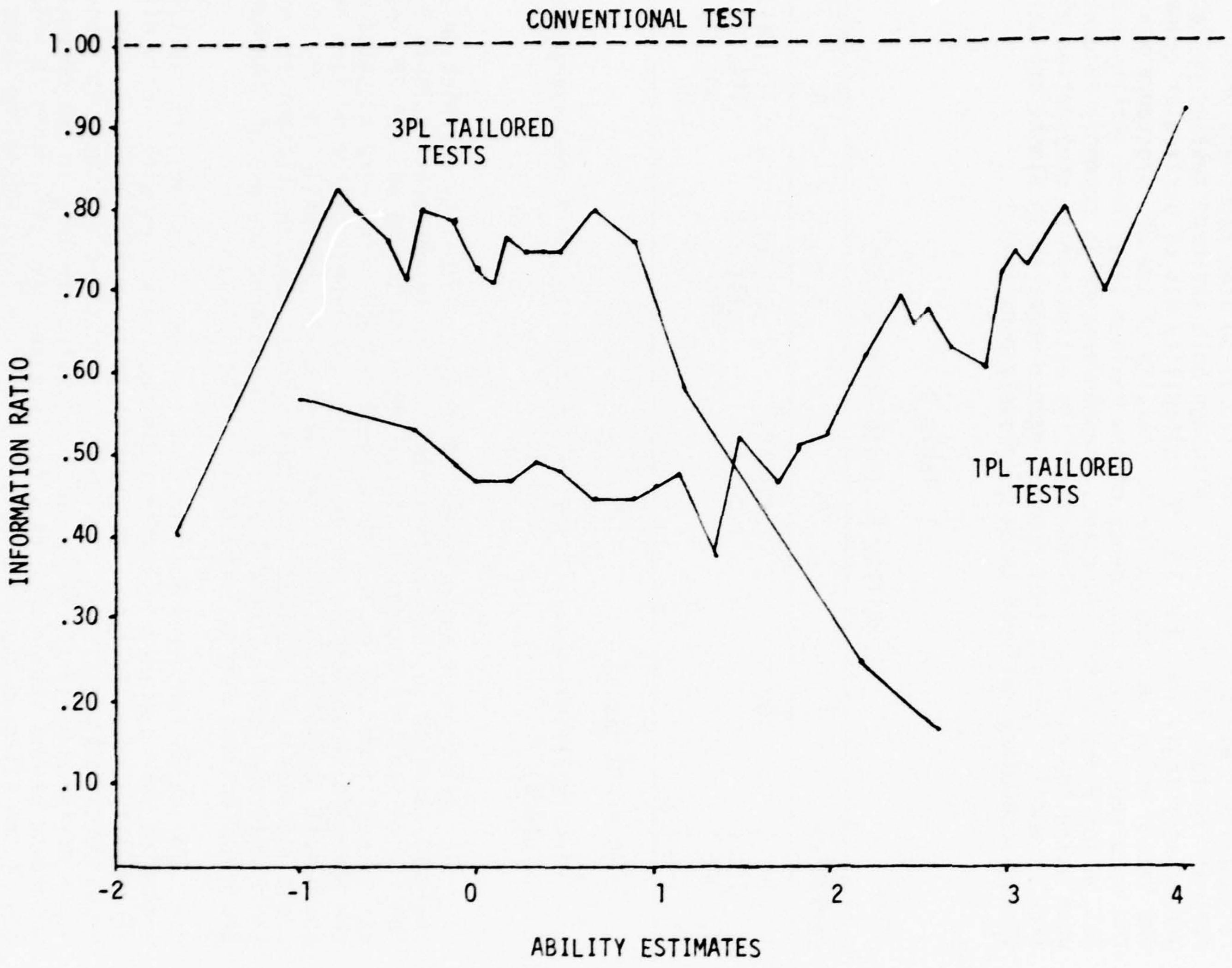
informative precisely in the ability range that encompassed most of the examinees.

Reliability

The correlation matrix in Table 3 reports the coefficients obtained from intercorrelating the ability estimates yielded by the two models

Figure 2

Relative Efficiency



in the tailored testing study. The .44 correlation between the ability estimates from the first 1PL test (1PL 1) and the second 1PL test (1PL 2) was the reliability coefficient for that procedure. This value, although by no means high, was significantly greater ($\underline{p}<.01$) than the .00 reliability coefficient obtained from the 3PL tailored testing procedure (3PL 1 vs. 3PL 2). Neither tailored testing procedure attained a reliability that approached the traditional 50 item paper-and-pencil form of the test (KR-20 = .74). Although both tailored testing reliabilities were disturbingly low, the 3PL .00 reliability was of particular concern. One factor which impacted on the reliability of the 3PL procedure was the occurrence of nonconvergence of the maximum likelihood ability estimation for 9 out of the 110 cases. Nonconvergence is commonly encountered when using the maximum likelihood ability estimation in conjunction with the 3PL model. (Recall that nonconvergence occurred in almost one-third of the vocabulary tailored tests previously mentioned.)

Table 3

Ability Estimate Correlations[a]

Variables	1	2	3	4
1. 1PL 1	1.00	.44(.46)[b]	.05(.31)	.12(.24)
2. 1PL 2		1.00	.11(.33)	.19(.13)
3. 3PL 1			1.00	.00(.12)
4. 3PL 2				1.00

[a] (n = 110 cases)

[b] (reliabilities when n = 101, due to deletion of 9 nonconvergence cases)

The deletion of these 9 cases from the reliability correlation analyses resulted in the coefficients shown in parentheses in Table 3. The 1PL reliability increased slightly from .44 to .46 and the 3PL reliability went from .00 to .12. When these reliabilities were adjusted with the Spearman-Brown formula to approximate the length of the 50 item paper-and pencil test, the 1PL coefficient went up to .68, while the 3PL coefficient increased to .25, both still being lower than the reliability of the traditional test. (Lord (1977) has questioned the use of Spearman-Brown corrections for tailored test reliabilities.)

To search further for sources of the low 3PL reliability, ability estimates were examined to locate individual examinees with widely differing 3PL ability scores from one test session to the next. Ten such cases were identified and studied in detail. These cases are shown in Table 4. A definite pattern emerged which reflected problems in the operating procedure of the tailored tests. All 10 cases were situations in which one of the tailored tests was only 3 or 4 items long, while the other was 20 items in length. The short test resulted when the examinee answered the initial and all the subsequent items correctly. Since there was never both a correct and incorrect response, no maximum likelihood ability estimate could be computed. Thus each successive item administered was more difficult by a fixed stepsize of about .693 on the ability scale. Ordinarily this would not be a problem with a good quality item pool. However, the achievement test item pool had only 28 out of 180 items above the zero point on the item difficulty scale. Moreover, the entry point into the pool had been set at +.50 or -.50. The result was that it was possible for an examinee to answer the first 3 or 4 tailored test items correctly by chance and "top out" of the item pool. When these cases of unreliable 3PL ability estimation were thrown out, the 3PL test reliability went up to .43. Obviously this was achieved only through substantial "massaging" of the data. It should be noted that the skewness of the item difficulties resulted mainly from the item linking procedures discussed earlier.

Table 4

Instances of Unreliable
3PL Ability Estimation

First Test Session		Second Test Session	
Number of Items	Ability Estimate	Number of Items	Ability Estimate
3	2.579	20	-.317
20	-.776	3	2.579
20	.152	3	2.579
20	.073	4	2.273
4	2.273	20	-.125
20	.010	3	2.579
20	-.270	4	2.273
20	.126	4	2.273
20	-.297	3	2.579
3	2.579	4	-1.700

Another problem with the 3PL tailored tests was that the item pool was functionally limited to only about 30 out of the 180 items. Since items were selected for administration based on the information function, only those items with relatively high item discrimination values were administered. The effect of this artificial restriction in the 3PL item pool was an overlap of more than 80% between the items administered from the first test session to the next. However, item repetition over tests was minimal for the 1PL tests. It seemed likely that common items across tests would favorably affect the 3PL reliability. However, partial correlation analyses in previous research indicated that the proportion of items in common had a negligible effect on the tailored test reliability.

Other Correlation Analyses

In Table 5 are listed the correlations computed between the tailored test estimated true scores and the scores on the three paper-and-pencil course exams, as well as a total exam score. Estimated true scores were calculated as simply the sum of the probabilities of correct responses to all the items in the pool for each examinee. In general, the correlations were relatively low. This was true even for the correlations between Exam 1 and the tailored tests, the case in which both types of tests covered the same content areas. The correlation between the first 3PL tailored test and Exam 1 was higher than the other tailored test correlations with Exam 1, but this could be due to chance alone. It should be noted that there were no major differences between the 1PL and 3PL tailored tests in terms of their correlations with the course exams. There does appear to be a substantial drop in the exam correlations from the first 3PL tailored test to the second 3PL test. However, this might have been due to the high number of cases of unreliable ability estimation for 3PL 2 (see Table 4).

Table 5

Correlations of Estimated True Scores[a]
With Traditional Course Exams[b]

Variables	1PL 1	1PL 2	3PL 1	3PL 2
Exam 1	.32	.38	.50	.29
Exam 2	.34	.25	.23	.14
Exam 3	.31	.18	.31	.25
Total Score	.56	.44	.48	.23

[a] (Calculated using the formula $\hat{\tau}_j(\theta) = \sum_{i=1}^{n} P_{ij}(\theta)$)

[b] (n = 101, since 9 nonconvergence cases were deleted from the analysis)

Descriptive Statistics

Table 6 presents some descriptive statistics for both test sessions of the two types of tailored tests. Since the administration of a maximum of 20 items was one stopping rule for the tailored tests, the values for the mean number of items administered indicate that most of the tests went the full distance. This result implied that an ample number of items was available in the item pool which had sufficient information for most of the examinees' ability estimates. The mean proportion of items correct reflected the overall low difficulty of the items for the majority of the students, since the mean proportion of items correct would have been expected to be .50 if the items were of exactly appropriate difficulty, assuming no guessing. The standard deviations of the ability estimates revealed that the scores yielded by the 3PL tailored tests had a restricted range compared to the 1PL tests, at least when the 10 unreliable cases were removed from the analyses.

Table 6

Tailored Test Descriptive Statistics[a]

Variable	One-Parameter Tailored Test		Three-Parameter Tailored Test	
	Session 1	Session 2	Session 1	Session 2
Mean # of items administered	19.56	19.72	19.18	18.10
Mean # of items correct	12.59	12.42	13.64	12.98
Mean proportion of items correct	.64	.63	.71	.72
Mean of ability estimates	1.74	1.75	.06	.18
S.D. of ability estimates	.87(.86)[b]	.80(.77)	.61(.27)	.79(.31)

[a] (n = 101, due to deletion of 9 nonconvergence cases)

[b] (n = 91, due to deletion of 10 cases with unreliable 3PL ability estimates)

Content Validity

As can be seen in Table 7, both the 1PL and 3PL item pools used for the tailored tests accurately reflected the weighting of the content areas in the paper-and-pencil course exam. Of course both item pools had identical content area breakdowns since the two pools contained the same items. A Chi-Square analysis indicated no lack of fit for the number of items in each content area of the pools compared to the corresponding number of items on the course exam. However, the number of items administered by content area for a systematic sample of 29 tailored tests showed significant lack of fit to both the item pools and the course exam. The fit of the 3PL tailored tests in terms of content validity was particularly bad, while the 1PL tests came fairly close to matching the content area weightings of the item pools and the course exam. It should be noted that no conscious attempt was made in the tailored testing operating program to require branching among the content areas. The object was to see if selecting items for administration on the basis of information would approximate the content area weightings of the item pools and the course exam. The multi-content nature of the course exam was demonstrated by numerous factor analyses indicating the presence of over 20 factors.

Attitude Scale Characteristics

The varimax rotated factor loading matrix that was obtained from a principal components analysis of the first administration of the attitude scale is shown in Table 8. A listing of the items on the attitude scale is in Appendix B. There were six factors present with eigenvalues greater than one, which accounted for 63% of the variance. The underlined values in the table indicate the highest factor loading for each item on the scale among the six factors. A subjective examination of the items loading on each factor resulted in the following factor labels:

Table 7

Test Items by Content Area for Course Exam, Item Pools, and Tailored Tests

Content Areas	Course Exam Items		Items in 1PL Pool		Items in 3PL Pool		Items in 29 1PL Tailored Tests		Items in 29 3PL Tailored Tests	
	Number	%	Number	%	Number	%	Number	%	Number	%
Anecdotal Records	5	10.0	17	9.4	17	9.4	49	9.1	57	10.4
Benavioral Objectives	5	10.0	18	10.0	18	10.0	56	10.3	28	5.1
Checklists	5	10.0	17	9.4	17	9.4	59	10.9	51	9.3
Peer Appraisals	2	4.0	7	3.9	7	3.9	13	2.4	0	0.0
Planning Tests	3	6.0	13	7.2	13	7.2	48	8.9	47	8.6
Rankings	3	6.0	11	6.1	11	6.1	26	4.8	10	1.8
Ratings	6	12.0	23	12.8	23	12.8	75	13.9	111	20.3
Selection Items	8	16.0	26	14.5	26	14.5	76	14.0	111	20.3
Self Report	2	4.0	7	3.9	7	3.9	32	5.9	45	8.2
Supply Items	5	10.0	19	10.6	19	10.6	62	11.5	26	4.7
Table of Specifica-tions	6	12.0	22	12.2	22	12.2	45	8.3	62	11.3
	50		180		180		541		548	

Note. Listed below are the Chi-Square values for several comparisons. The critical values for rejection of adequate fit is $\chi^2(10) > 18.31$ at $\alpha = .05$.

1. Course exam items vs. items in 1PL pool, $\chi^2 = .9978$
2. Course exam items vs. items administered by 1PL tailored tests, $\chi^2 = 28.245$
3. Items in 1PL pool vs. items administered by 1PL tailored tests, $\chi^2 = 21.383$
4. Course exam items vs. items in 3PL pool, $\chi^2 = .9978$
5. Course exam items vs. items administered by 3PL tailored tests, $\chi^2 = 134.341$
6. Items in 3PL pool vs. items administered by 3PL tailored tests, $\chi^2 = 133.448$

factor I - perceived test performance
factor II - anxiety
factor III - cathode ray tube (CRT) characteristics
factor IV - motivation/test satisfaction
factor V - item easiness
factor VI - time pressure

Table 8
Principal Components Analysis:
Varimax Rotated Factor Pattern for
First Attitude Scale Administration

Item No.	I	II	III	IV	V	VI
1	.23	-.60	-.33	.05	-.08	-.08
2	.03	-.32	-.02	.01	-.02	-.68
3	.43	-.16	.24	.28	-.41	-.19
4	.01	-.67	.06	-.15	.23	-.05
5	.15	.20	.64	-.38	.15	.03
6	.78	-.05	.01	-.09	-.15	.05
7	.52	-.20	-.07	-.55	.11	-.33
8	.16	-.25	.62	.11	.02	-.17
9	-.08	.00	.10	.04	.90	-.07
10	.26	-.71	.14	.13	.06	.16
11	.06	.18	.37	-.02	.08	-.66
12	.66	-.25	.19	-.30	.21	.03
13	.06	-.72	.32	-.16	-.19	-.25
14	.82	-.05	.04	-.12	-.10	-.08
15	.06	-.64	.10	-.35	-.20	-.32
16	-.15	-.42	.59	-.12	-.16	-.36
17	.47	-.20	.02	-.53	.16	-.36
18	.19	-.01	-.12	-.62	.12	.03
19	.04	-.27	.42	-.61	.02	.06
20	.03	-.00	.17	-.71	.20	-.00

Note. The underlined values indicate the highest loading of an item on a factor. Broken underlines indicate other high loadings.

The data available from the second attitude scale administration were also submitted to a principal components analysis with a varimax rotation. Again six factors were present with eigenvalues greater than one, accounting for 65% of the variance. The rotated factor loading matrix is presented in Table 9. The numbering of the factors changed somewhat and a few of the items switched factors. However, the general pattern of components was the same as the results from the first scale administration. The labeled factors are listed below:

factor I - anxiety/time pressure
factor II - motivation
factor III - perceived test performance
factor IV - item easiness
factor V - test satisfaction
factor VI - CRT characteristics

Table 9

Principal Components Analysis:
Varimax Rotated Factor Pattern for
Second Attitude Scale Administration

Item No.	I	II	III	IV	V	VI
1	.43	-.41	-.23	-.34	.03	.05
2	.51	-.10	.14	-.05	-.50	.26
3	.33	-.18	-.53	.04	-.05	.09
4	.81	.12	-.07	-.03	-.24	-.09
5	-.03	.78	.01	.11	-.17	-.11
6	.08	.10	-.73	-.09	-.10	.03
7	.12	.16	-.34	.17	-.80	.10
8	.16	-.05	-.25	.07	.03	.70
9	-.01	.03	.22	.75	-.12	-.03
10	.62	.00	-.13	.07	.09	.25
11	.20	.18	-.19	.61	-.09	.23
12	.10	.18	-.61	-.07	-.52	-.14
13	.82	.17	-.09	.16	-.10	.07
14	.14	.15	-.78	.00	-.33	.05
15	.76	.03	-.24	-.01	-.11	.15
16	.14	-.01	.16	.01	-.14	.83
17	.12	-.01	-.33	.13	-.85	.01
18	.06	.51	.09	-.43	-.38	.15
19	.07	.71	-.29	.07	-.06	.19
20	.17	.76	-.06	.03	.13	-.14

Note. The underlined values indicate the highest loading of an item on a factor. Broken underlines indicate other high loadings.

As previously mentioned, several other exploratory factor analyses were run on the attitude scale data. These included both alpha factor analysis and image covariance analysis, each with varimax, oblimax, and binormamin rotations. Although the results were quite similar in all cases, the image covariance analysis with varimax rotation was judged to be the most satisfactory solution. The factor loading matrix resulting from the first attitude scale administration is presented in Table 10. The labeled factors are listed below:

factor I - anxiety/time pressure
factor II - perceived test performance
factor III - motivation
factor IV - CRT characteristics
factor V - test satisfaction
factor VI - item easiness

Table 10

Image Covariance Factor Analysis:
Varimax Rotated Factor Pattern for First
Attitude Scale Administration

Item No.	I	II	III	IV	V	VI
1	.53	.23	-.10	-.06	.22	-.01
2	.40	.07	.02	.33	.24	.00
3	.31	.43	.02	.33	-.07	-.13
4	.60	.09	.23	.07	.03	.23
5	-.04	.18	.55	.35	-.11	.23
6	.17	.69	.24	.09	.07	-.04
7	.24	.39	.43	.09	.49	.08
8	.26	.17	.07	.55	-.08	.16
9	.07	-.01	.18	.14	.06	.50
10	.65	.28	.07	.08	-.08	.13
11	-.02	.09	.11	.52	.17	.11
12	.19	.59	.24	.07	.21	.24
13	.73	.13	.26	.35	.05	-.01
14	.14	.76	.18	.10	.12	-.02
15	.66	.13	.36	.27	.19	-.11
16	.47	-.04	.29	.64	-.01	-.01
17	.18	.40	.35	.17	.53	.15
18	.07	.18	.50	-.07	.20	.10
19	.33	.19	.60	.26	.03	.20
20	.15	.16	.60	.16	.09	.03

Note. The underlined values indicate the highest loading of an item on a factor. Broken underlines indicate other high loadings.

It should be noted that the factor analysis results obtained from the attitude scale administrations were similar to those obtained in a previous study (Koch and Reckase, 1978). However, one difference was that anxiety and time pressure items formerly loaded on separate factors, while in the present study they usually formed one factor. Also, item 9 on the attitude scale now loaded on its own factor, while previously it loaded with other perceived test performance items. The only real difference in the attitude scale itself was that the former items referred to a vocabulary test while the present items referred to the course content areas.

Two types of reliability measures were calculated for the attitude scale. First, a test-retest reliability coefficient was computed between the sets of total attitude scores for the two tailored testing sessions. The total attitude scores were obtained for each examinee by summation of the scores on the 20 individual items. A value of r = .71 was obtained, based on 104 cases with attitude data for both sessions.

Table 11

Coefficient Alpha Reliabilities for the Attitude Scale Factors and Total Scale

Factor Labels	Items	Coeff. α Session 1	Coeff. α Session 2
Anxiety/Time Pressure	1, 2, 4 10, 13, 15	.73	.74
Perceived Test Performance	3, 6, 12, 14	.69	.71
Motivation	5, 18, 19, 20	.59	.67
CRT Characteristics	8, 11, 16	.58	.47
Test Satisfaction	7, 17	.82	.92
Total Scale	all 20 items[a]	.81	.80

[a] Item 9 loaded on its own factor, so no Coefficient α could be calculated for it alone.

Secondly, coefficient alpha reliabilities were calculated for the factors of the scale as well as for the total scale itself. The results are shown in Table 11. In general the coefficients were fairly high for all of the individual factors as well as for the total scale. However, item 9 loaded on its own factor during the factor analysis, so no coefficient alpha could be figured for that particular item. In previous research using the attitude scale, item 9 had a low discrimination index with respect to total score. Since the present study duplicated that finding, it was decided to rewrite item 9 for future administrations of the scale. The item discrimination indices were calculated by correlating individual item scores with total scores for each examinee. The results are shown in Table 12. Again, all of the item discrimination values were reasonably high except for item 9.

Attitude Scale Results

The responses obtained from the administration of the scale are summarized in Tables 13-17. Response percentages to the five categories for each item are shown, with the items grouped together according to the factors measured by the attitude scale.

The results in Table 13 indicate that the examinees generally felt very little time pressure on the tailored tests. Moreover the students expressed that they experienced little stress or nervousness in regard to the tailored tests. For the perceived test performance factor results shown in Table 14, the students' responses were quite evenly split as to their judgments of their own performances. However, item 3 showed that most of the examinees disagreed with the statement that the test items were too difficult. The responses summarized in Table 15 indicate

Table 12

Discrimination Indexes for Attitude Scale Items for Two Test Sessions

Item No.	Session 1	Session 2
1	.38	.26
2	.45	.53
3	.33	.45
4	.50	.62
5	.36	.30
6	.41	.44
7	.68	.71
8	.48	.40
9	.07	.08
10	.46	.45
11	.35	.46
12	.65	.61
13	.67	.62
14	.51	.63
15	.64	.63
16	.54	.33
17	.68	.65
18	.35	.38
19	.53	.48
20	.38	.26

that the examinees reported that they were highly motivated to do well on the tailored tests. However, a substantial percentage of students felt that they could have done better on the test if they had tried harder.

Table 16 summarizes the responses for the cathode ray tube (CRT characteristics. Most of the examinees found the screen easy to read and experienced little eye discomfort. However, there was a split vote on whether or not the pace of the computer was too slow during the administration of the tailored tests. Finally, Table 17 shows the results for the test satisfaction factor. Opinion was evenly divided on the issue of whether the tailored tests did a good job of measuring their abilities. Many of the students did not feel that their performances on the tests reflected their "true" knowledge of the course material.

The results of the MANOVA to determine if any changes occurred in attitudes across test sessions were non-significant. The implication was that student attitudes toward the various aspects of the tailored testing situation did not differ from one test session to the next.

Table 13

Response Percentages for the Anxiety/Time Pressure Factor for Items and Alternatives over Both Sessions

1. During the test I was worried about how well I was doing.

	session 1	session 2
strongly agree	7	3
agree	40	28
neutral	17	22
disagree	28	38
strongly disagree	8	9

2. I felt less time pressure while taking this test than while taking conventional tests.

	session 1	session 2
strongly agree	25	17
agree	41	48
neutral	14	10
disagree	16	23
strongly disagree	4	2

4. The computer terminal made me feel that I had to answer the items as quickly as possible.

	session 1	session 2
strongly agree	2	1
agree	8	3
neutral	9	9
disagree	55	62
strongly disagree	26	25

10. I was nervous about coming here to take this test.

	session 1	session 2
strongly agree	0	1
agree	10	0
neutral	11	5
disagree	54	67
strongly disagree	25	27

13. The computer terminal made me nervous.

	session 1	session 2
strongly agree	2	0
agree	6	5
neutral	6	6
disagree	61	66
strongly disagree	25	23

15. I felt considerable stress while taking the test.

	session 1	session 2
strongly disagree	26	17
disagree	58	72
neutral	9	7
agree	7	4
strongly agree	0	0

Table 14

Response Percentages for the Perceived Difficulty Factor for Items and Alternatives over Both Sessions

3. I felt that many of the items were too difficult for me.

	session 1	session 2
strongly disagree	1	4
disagree	58	55
neutral	31	33
agree	10	7
strongly agree	0	1

6. I think I did well on the test compared to other people.

	session 1	session 2
strongly agree	0	0
agree	23	27
neutral	65	64
disagree	12	9
strongly disagree	0	0

12. I feel that I did as well on this test as on other tests I've taken.

	session 1	session 2
strongly agree	7	2
agree	33	39
neutral	22	27
disagree	31	27
strongly disagree	7	5

14. I felt confident that I did well on the test.

	session 1	session 2
strongly disagree	2	1
disagree	24	18
neutral	43	52
agree	30	29
strongly agree	1	0

Table 15

Response Percentages for the Motivation Factor for Items and Alternatives over Both Sessions

5. I didn't care very much about how well I did on the test.

	session 1	session 2
strongly disagree	12	9
disagree	54	57
neutral	25	14
agree	7	19
strongly agree	2	1

18. I think I could have done better on the test if I had tried harder.

	session 1	session 2
strongly disagree	3	3
disagree	29	31
neutral	28	25
agree	35	37
strongly agree	5	4

19. I was careful to try to select the best answer to each question.

	session 1	session 2
strongly disagree	0	0
disagree	1	4
neutral	7	7
agree	67	72
strongly agree	25	17

20. I tried to finish the test quickly just to receive my 5 points credit.

	session 1	session 2
strongly agree	2	0
agree	2	2
neutral	10	12
disagree	58	70
strongly disagree	28	16

Table 16

Response Percentages for the CRT Characteristics Factor for Items and Alternatives over Both Sessions

	session 1	session 2
8. My eyes were uncomfortable when viewing the screen.		
strongly agree	5	2
agree	21	18
neutral	10	8
disagree	49	60
strongly disagree	15	12
11. The pace of the computer was so slow that it made me impatient.		
strongly disagree	6	2
disagree	37	39
neutral	19	16
agree	30	35
strongly agree	8	8
16. It was easy to read the words and questions on the screen.		
strongly agree	30	17
agree	55	67
neutral	7	8
disagree	8	5
strongly disagree	0	3

Table 17

Response Percentages for the Test Satisfaction Factor for Items and Alternatives over Both Sessions

7. I felt that my performance on this test reflected my true knowledge of A140.

	session 1	session 2
strongly disagree	6	3
disagree	35	40
neutral	33	27
agree	26	30
strongly agree	0	0

17. I felt that the test did a good job of measuring my ability in A140.

	session 1	session 2
strongly agree	0	1
agree	21	28
neutral	43	31
disagree	32	36
strongly disagree	4	4

Discussion

Goodness of Fit

The superior fit of the observed responses to those predicted by the 3PL model was expected based on previous research (Koch and Reckase, 1978; Reckase, 1977). It was not surprising that a model with three item parameters was able to fit observed response data better than a model with only one item parameter. Since the MSD values reflected an average fit across the response string for an examinee, the implication can be made that the 3PL tailored tests demonstrated better "person fit" than the 1PL tests.

Information Function Analyses

The results of the relative efficiency comparisons shown in Figure 1 clearly demonstrated the inadequacy of both the 1PL and the 3PL tailored achievement tests compared to the traditional paper-and-pencil achievement test. This result was contrary to the findings of previous tailored testing research with vocabulary ability tests. In the latter case, 3PL tailored tests averaging 19 items were more than twice as informative as the 30 item conventional vocabulary test at certain points on the ability scale. Since the achievement tailored tests averaged only about 20 items in length compared to the 50 item course exam, a drop was expected in the tailored test relative efficiency. This was predicated since total test information is just the sum of the item information. However, it was not expected that the 1PL tailored tests would be only about half as informative and the 3PL tailored tests only about 80% as informative as the conventional course exam. No conclusive explanation could be identified for this result. Perhaps the item parameter linking procedures were at fault.

Certainly it was true that the tailored tests had more information on a per item basis. However, that is beside the point. Part of the merit of tailored tests is that a shortened test may be as informative about an examinee's ability as the conventional full length test. This is accomplished through more accurate measurement by the administration of only the appropriate test items. Clearly, further research is required. A final curious result was that the 3PL tailored tests were more informative than the 1PL tests in the ability range where most of the examinees were concentrated, even though the 1PL tailored tests were significantly more reliable.

Reliability

The reliability results provided another setback for the tailored testing procedures. As has been mentioned earlier, the previous vocabulary tailored testing study yielded adequately high reliabilities for both the 1PL and the 3PL procedures, the values being $r = .61$ and $r = .77$, respectively. But the tailored achievement test reliabilities did not even approach the course exam reliability. Moreover, the 3PL procedure had zero reliability, for which several contributing factors were identified.

One major problem was that the item parameter linkings resulted in a somewhat skewed and shifted distribution of the 3PL difficulty parameters so that only about 30 out of 180 items were above the zero point on the scale. This outcome, in combination with the tailored testing operational procedures of the ±.50 entry point and the fixed stepsize, resulted in unreliable tests for numerous examinees. In hindsight, the entry point into the item pool should have been shifted downward on the difficulty scale so that approximately an equal number of items were above and below the starting point. In that situation, examinees who were able to answer the first few items correctly would not have been able to "top out" of the item pool.

Nonconvergence of maximum likelihood ability estimation was another problem with the 3PL tailored tests. When the very large number of nonconvergence cases was observed in the previous vocabulary study, the hypothesis was forwarded that excessively difficult items were the cause, yielding long strings of incorrect responses. In such a case no reasonable maximum likelihood ability estimation could be calculated since the likelihood function approached a uniform distribution with the ordinate at the guessing level. Since the achievement tailored tests were based on the examinee's regular course material over which they had been previously tested, the nonconvergence problem was reduced somewhat, with only 9 out of 110 failures to converge. Several approaches are currently being studied to resolve the nonconvergence problem, including the alternative of substituting Bayesian ability estimation in place of maximum likelihood.

Since neither of the problems discussed immediately above applied to the 1PL tailored tests, another explanation must be found for the low reliability of that procedure. The most obvious candidate is the multidimensionality of the test. Since the principal components analysis of the regular course Exam 1 indicated the presence of 20 factors with eigenvalues greater than one, it was obvious that the unidimensional assumption of the latent trait models had been violated. Therefore, the low 1PL reliability could have simply been a result of the violation of that assumption and its negative effect on item calibration. Of course, the same argument would apply to the 3PL tailored tests. If indeed future research shows that the latent trait models are not robust with respect to the violation of the unidimensionality assumption, then each content area of achievement tests will have to be identified and calibrated separately. In addition, intricate branching schemes will have to be devised so that the tailored tests can provide ability estimates for each content area. Scoring would then become a problem in terms of weighting the content areas. If the content areas were correlated somewhat, it might be possible to use regression methods to predict the appropriate entry point into a new content area, given an ability estimate on the previous content area (Brown and Weiss, 1977).

Content Validity

The content validity results demonstrated that, even though the item pools may reflect proportionate content area weightings to a conventional test, the tailored tests using the item pools did not necessarily reflect the same weightings. For the 1PL procedure this result was somewhat of a surprise, since the 1PL tailored tests utilized most of the items in the pool. In such cases, the tailored tests should have performed similarly to a random sampling process from the item pools. However, for the 3PL tailored tests, only the most discriminating items were administered, regardless of content areas, since items were selected for administration on the basis of the information function. Item discrimination values do not come into play for the 1PL procedure since they are all assumed to be one. Perhaps if a larger sample than 29 tailored tests had been analyzed, the 1PL procedure would have achieved adequate content validity.

In contrast, 3PL tailored testing procedures will undoubtedly require branching schemes from one content area to another in order to insure adequate weighting of all the content areas. It is interesting to note that the correlation between the first 3PL tailored test estimated true scores and the scores on course Exam 1 was quite high, despite the poor content validity of the 3PL tailored tests (see Table 5). No explanation for this anomaly is available. In this regard, content validity might be more appropriately measured in terms of amount of information or precision of measurement in each content area rather than just number of items.

Attitude Scale

The attitude scale to measure the students' attitudes toward tailored testing was found to have fairly high reliability. The test-retest reliability coefficient was r = .70, and the coefficient α reliability was r = .80. In addition, all of the items on the scale correlated higher than +.30 with the total score except for item 9, which will be revised for future administrations. The factor analysis results for both testing sessions were found to be quite similar, with five main factors represented on the attitude scale.

The results from the attitude scale response data were generally favorable toward tailored testing. For instance, the majority of the examinees said they experienced very little time pressure during the tests. Moreover, few students reported any nervousness or stress in connection with the tailored test or the computer terminal. Most of the examinees reported that the CRT screen was comfortable to view and that it was easy to read the test items. A majority of the students' responses reflected fairly high levels of motivation to do well on the tailored tests and to carefully select answers. However, the opinions of the students were divided on whether they could have done better on the tests if they had tried harder. It should be noted that inconsistencies in the scale results may call into question the care with which the students filled out their attitude surveys.

The responses to the items for the perceived test performance factor were about as expected. That is, responses were evenly divided when the examinees attempted to judge how well they performed on the tailored tests. This result was expected since the aim of the tailored tests was to administer items of appropriate difficulty for each examinee. Then too, no feedback was provided as to the correctness of responses. On the more negative side, attitude responses were quite evenly divided on the issue of whether or not the tailored tests did a good job of measuring the students' ability levels or their "true" knowledge of the course material.

Previous attitude research had failed to show any significant correlations between the attitudes of the students toward the tailored tests and their performance on the tests. The presenty study again yielded no evidence of any linear relationship in this regard. Even though such variables as motivation and anxiety levels might be expected to interact with test performance, the present research provided no evidence to support such effects.

Summary and Conclusion

The results of applying tailored testing procedures to the measurement of unidimensional vocabulary ability were generally satisfactory. Reliabilities and information were comparable to or better than the conventional test for both the 1PL and 3PL tests. However, tailored testing applied to multidimensional achievement measurement presented many difficulties. Both the 1PL and 3PL procedures were inadequate with regard to reliability, test information, and content validity. Possible causes were the small sample sizes used to calibrate the tests, resulting in unstable item parameter estimates; a compounding of the instability of the parameter estimates during linking procedures; poor selection of entry points into the item pools; the possibility that latent trait models may not be robust with respect to violation of the unidimensionality assumption by multi-content achievement tests; and the nonconvergence of the 3PL tailored tests when using maximum likelihood ability estimation.

One way to look at the present study is to view it as an example of mistakes not to make in tailored achievement testing. From perhaps a more reasonable perspective, the study illustrates that very little can be taken for granted in setting up tailored testing procedures. Rather, one must carefully make decisions about the operational procedures, while considering the effects that such decisions might have. A great deal more research must be conducted to determine optimal levels of the various components that control tailored testing procedures. A study by Patience and Reckase (1979) is an important step in this direction.

REFERENCES

Bejar, I. I., Weiss, D. J. and Kingsbury, G. G. *Calibration of an item pool for the adaptive measurement of achievement* (Research Report 77-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1977.

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

Brown, J. M. and Weiss, D. J. *An adaptive testing strategy for achievement test batteries* (Research Report 77-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1977.

Koch, W. R. and Reckase, M. D. *A live tailored testing comparison study of the one- and three-parameter logistic models* (Research Report 78-1). Columbia: University of Missouri, Department of Educational Psychology, 1978.

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

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. Discussion: session 2. In D. J. Weiss (Ed.), *Proceedings of the 1977 computerized adaptive testing conference*. Minneapolis: University of Minnesota, Department of Psychology, 1977.

Lord, F. M. Personal communication, June, 1979.

Lord, F. M. and Novick, M. R. *Statistical theories of mental test scores*. Reading, Massachusetts: Addison-Wesley, 1968.

Patience, W. M. Description of components in tailored testing. *Behavior Research Methods and Instrumentation*, 1977, *9*, 153-157.

Patience, W. M. and Reckase, M. D. *Operational characteristics of a Rasch model tailored testing procedure when program parameters and item pool attributes are varied*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, April, 1979.

Rasch, G. *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research, 1960.

Reckase, M. D. *Ability estimation and item calibration using the one- and three-parameter logistic models: a comparative study*, (Research Report 77-1). Columbia: University of Missouri, Department of Educational Psychology, 1977.

Reckase, M. D. *Item pool construction for use with latent trait models.* Paper presented at the Annual Meeting for the American Educational Research Association, San Francisco, April, 1979.

Urry, V. W. *Ancillary estimators for the item parameters of mental test models.* Paper presented at the Annual Meeting of the American Psychological Association, Chicago, August, 1975.

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

Wood, R. L., Wingersky, M. S., and Lord, F. M. *LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters* (ETS Research Memorandum RM-76-6). Princeton, New Jersey: Educational Testing Service, June, 1976.

Wright, B. D. Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 1977, *14*, 97-116.

Wright, B. D. and Panchapakesan, N. A procedure for sample-free item analysis. *Educational and Psychological Measurement*, 1969, *29*, 23-48.

APPENDIX A - 1

Table A - 1

Classroom Achievement Tests Calibrated
for Tailored Testing Usage

One-parameter Calibration

Date	Sample Size
10-72	258
12-72	170
2-73	305
4-73	224
9-74	205
9-75	203
4-76	187
9-76	177
11-76	96
2-77 & 4-77	314
9-77 & 10-77	202

Three-parameter Calibration

Date	Sample Size
4-76	187
9-76	177
11-76	97
2-77 & 4-77	314
9-77 & 10-77	202

APPENDIX A - 2

Table A - 2

Table of Specifications for Exam I

Content Areas	Knowledge of Terms and Techniques	Application of Techniques	Analysis, Synthesis, and Evaluation of Techniques	Totals
Planning the Test	1	1	1	3
Behavioral Objectives	1	2	2	5
Table of Specifications	2	2	2	6
Anecdotal Records	1	2	2	5
Rating Scales	2	2	2	6
Checklists	1	2	2	5
Rankings	1	1	1	3
Peer Appraisals	1	1		2
Self Reports	1	1		2
Selection Items	2	3	3	8
Supply Items	1	2	2	5
	14	19	17	50

APPENDIX B

Attitude Survey Administered after each Tailored Test Session

Please circle the response to each statement below which most nearly reflects your feelings or attitude.

1. During the test I was worried about how well I was doing.

strongly				strongly
agree	agree	neutral	disagree	disagree

2. I felt less time pressure while taking this test than while taking conventional tests.

strongly				strongly
agree	agree	neutral	disagree	disagree

3. I felt that many of the items were too difficult for me.

strongly				strongly
disagree	disagree	neutral	agree	agree

4. The computer terminal made me feel that I had to answer the items as quickly as possible.

strongly				strongly
agree	agree	neutral	disagree	disagree

5. I didn't care very much about how well I did on the test.

strongly				strongly
disagree	disagree	neutral	agree	agree

6. I think I did well on the test compared to other people.

strongly				strongly
agree	agree	neutral	disagree	disagree

7. I felt that my performance on this test reflected my true knowledge of A140.

strongly				strongly
disagree	disagree	neutral	agree	agree

8. My eyes were uncomfortable when viewing the screen.

strongly				strongly
agree	agree	neutral	disagree	disagree

9. I felt that many of the items on the test were too easy.

strongly				strongly
disagree	disagree	netural	agree	agree

10. I was nervous about coming here to take this test.

strongly agree	agree	neutral	disagree	strongly disagree

11. The pace of the computer was so slow that it made me impatient.

strongly disagree	disagree	neutral	agree	strongly agree

12. I feel that I did as well on this test as on other tests I've taken.

strongly agree	agree	neutral	disagree	strongly disagree

13. The computer terminal made me nervous.

strongly agree	agree	neutral	disagree	strongly disagree

14. I felt confident that I did well on the test.

strongly disagree	disagree	netural	agree	strongly agree

15. I felt considerable stress while taking the test.

strongly disagree	disagree	neutral	agree	strongly agree

16. It was easy to read the words and questions on the screen.

strongly agree	agree	neutral	disagree	strongly disagree

17. I felt that the test did a good job of measuring my ability in A140.

strongly agree	agree	neutral	disagree	strongly disagree

18. I think I could have done better on the test if I had tried harder.

strongly disagree	disagree	neutral	agree	strongly agree

19. I was careful to try to select the best answer to each question.

strongly disagree	disagree	neutral	agree	strongly agree

20. I tried to finish the test quickly just to receive my 5 points credit.

strongly agree	agree	neutral	disagree	strongly disagree

DISTRIBUTION LIST

Navy

1 Dr. Ed Aiken
Navy Personnel R & D Center
San Diego, CA 92152

1 Dr. Jack R. Borsting
Provost & Academic Dean
U.S. Naval Postgraduate School
Monterey, CA 93940

1 Dr. Robert Breaux
Code N-71
NAVTRAEQUIPCEN
Orlando, FL 32813

1 Mr. Maurice Callahan
Pers 23a
Bureau of Naval Personnel
Washington, DC 20370

1 Dr. Pat Federico
Navy Personnel R & D Center
San Diego, CA 92152

1 Dr. Paul Foley
Navy Personnel R & D Center
San Diego, CA 92152

1 Dr. John Ford
Navy Personnel R & D Center
San Diego, CA 92152

1 Capt. D. M. Gragg, MC, USN
Head, Section on Medical Ed.
Uniformed Services Univ. of the
Health Sciences
6917 Arlington Road
Bethesda, MD 20014

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

1 Dr. Leonard Kroeker
Navy Personnel R & D Center
San Diego, CA 92152

1 Chairman, Leadership & Law Dept.
Div. of Professional Development
U.S. Naval Academy
Annapolis, MD 21402

Navy

1 Dr. William L. Maloy
Principal Civilian Advisor for
Education and Training
Naval Training Command
Code 00A
Pensacola, FL 32508

1 Capt. Richard L. Martin
USS Francis Marion (LPA-Z49)
FPO New York, NY 09501

1 Dr. James McBride
Code 301
Navy Personnel R & D Center
San Diego, CA 92152

2 Dr. James McGrath
Navy Personnel R & D Center
Code 306
San Diego, CA 92152

1 Dr. William Montague
LRDC
University of Pittsburgh
3939 O'Hara Street
Pittsburgh, PA 15213

1 Library
Navy Personnel R & D Center
San Diego, CA 92152

6 Commanding Officer
Naval Research Laboratory
Code 2627
Washington, DC 20390

1 Office of Civilian Personnel
(Code 26)
Dept. of the Navy
Washington, DC 20390

1 John Olsen
Chief of Naval Education &
Training Support
Pensacola, FL 32509

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

Navy

1 Psychologist
ONR Branch Office
536 S. Clark Street
Chicago, IL 60605

1 Code 436
Office of Naval Research
Arlington, VA 22217

1 Office of Naval Research
Code 437
800 N. Quincy Street
Arlington, VA 22217

5 Personnel & Training Research
Programs (Code 458)
Office of Naval Research
Arlington, VA 22217

1 Psychologist
Office of Naval Research Branch
223 Old Marylebone Road
London, NW, 15th England

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

1 Scientific Director
Office of Naval Research
Scientific Liaison Group/Tokyo
American Embassy
APO San Francisco, CA 96503

1 Office of the Chief of Naval
Operations
Research, Development, & Studies
Branch (OP-102)
Washington, DC 20350

1 Scientific Advisor to the Chief
of Naval Personnel (Pers-Or)
Naval Bureau of Personnel
Room 4410, Arlington Annex
Washington, DC 20370

1 Dr. Richard A. Pollak
Academic Computing Center
U.S. Naval Academy
Annapolis, MD 21402

Navy

1 Mr. Arnold Rubenstein
Naval Personnel Support Tech.
Naval Material Command (08T244)
Room 1044, Crystal Plaza #5
2221 Jefferson Davis Highway
Arlington, VA 20360

1 A. A. Sjoholm
Tech. Support, Code 201
Navy Personnel R & D Center
San Diego, CA 92152

1 Mr. Robert Smith
Office of Chief of Naval Oper.
OP-987E
Washington, DC 20350

1 Dr. Alfred E. Smode
Training Analysis & Evaluation
Group (TAEG)
Dept. of the Navy
Orlando, FL 32813

1 Dr. Richard Sorensen
Navy Personnel R & D Center
San Diego, CA 92152

CDR Charles J. Theisen, Jr. MSC,
USN
Head Human Factors Engineering Div.
Naval Air Development Center
Warminster, PA 18974

1 W. Gary Thomson
Naval Ocean Systems Center
Code 7132
San Diego, CA 92152

1 Dr. Ronald Weitzman
Dept. of Administrative Sciences
U.S. Naval Postgraduate School
Monterey, CA 93940

1 Dr. Martin F. Wiskoff
Navy Personnel R & D Center
San Diego, CA 92152

Army

1 Technical Director
U.S. Army Research Institute for
the Behavioral & Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

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

1 Dr. Ralph Canter
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Ralph Dusek
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Myron Fischl
U.S. Army Research Institute for
the Social & Behavioral Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Harold F. O'Neil, Jr.
ATTN: PERI-OK
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Robert Ross
U.S. Army Research Institute for
the Social & Behavioral Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Director, Training Development
U.S. Army Administration Center
ATTN: Dr. Sherrill
Ft. Benjamin Harrison, IN 46218

1 Dr. Frederick Steinheiser
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Joseph Ward
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Air Force

1 Air Force Human Resources Lab
AFHRL/PED
Brooks AFB, TX 78235

1 Air University Library
AUL/LSU 76/443
Maxwell AFB, AL 36112

1 Dr. Philip De Leo
AFHRL/TT
Lowry AFB, CO 80230

1 Dr. G. A. Eckstrand
AFHRL/AS
Wright-Patterson AFB, OH 45433

1 CDR. Mercer
CNET Liaison Officer
AFHRL/Flying Training Div.
Williams AFB, AZ 85224

1 Dr. Ross L. Morgan (AGHRL/ASR)
Wright-Patterson AFB
Ohio 45433

1 Dr. Roger Pennell
AFHRL/TT
Lowry AFB, CO 80230

1 Personnel Analysis Division
HQ USAF/DPXXA
Washington, DC 20330

1 Research Branch
AFMPC/DRMYP
Randolph AFB, TX 78148

1 Dr. Malcolm Ree
AFHRL/PED
Brooks AFB, TX 78235

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

1 Jack A. Thorpe, Capt., USAF
Program Manager
Life Scien-es Directorate
AFOSR
Bolling AFB, DC 20332

Air Force

1 Brian K. Waters, LCOL, USAF
Air University
Maxwell AFB
Montgomery, AL 36112

Coast Guard

1 Mr. Richard Lanterman
Psychological Research (G-P-1/62)
U.S. Coast Guard HQ
Washington, DC 20590

1 Dr. Thomas Warm
U.S. Coast Guard Institute
P.O. Substation 18
Oklahoma City, OK 73169

Marines

1 Director, Office of Man. Util.
HQ, Marine Corps (MPU)
BCB, Bldg. 2009
Quantico, VA 22134

1 MCDEC
Quantico Marine Corps Base
Quantico, VA 22134

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

Other DoD

12 Defense Documentation Center
Cameron Station, Bldg. 5
Alexandria, VA 22314
ATTN: TC

1 Dr. Dexter Fletcher
Advanced Research Projects Agency
1400 Wilson Blvd.
Arlington, VA 22209

1 Military Assistant for Training
& Personnel Technology
Office of the Under Secretary of
Defense for Research & Engineering
Room 3D129, The Pentagon
Washington, DC 20301

1 Major Wayne Sellman, USAF
Office of the Assistant Secretary
of Defense (MRA&L)
3B930 The Pentagon
Washington, DC 20301

Civil Govt.

1 Dr. Susan Chipman
Basic Skills Program
National Institute of Education
1200 19th Street NW
Washington, DC 20208

1 Dr. William Gorham, Director
Personnel R & D Center
U.S. Civil Service Commission
1900 E. Street NW
Washington, DC 20415

1 Dr. Joseph I. Lipson
Division of Science Education
Room W-638
National Science Foundation
Washington, DC 20550

1 Dr. John Mays
National Institute of Education
1200 19th Street NW
Washington, DC 20208

1 Dr. Arthur Melmed
National Institute of Education
1200 19th Street NW
Washington, DC 20208

1 Dr. Andrew R. Molnar
Science Education Dev & Research
National Science Foundation
Washington, DC 20550

1 Dr. Lalitha P. Sanathanan
Environmental Impact Studies Div.
Argonne National Laboratory
9700 S. Cass Avenue

1 Dr. Jeffrey Schiller
National Institute of Education
1200 19th Street NW
Washington, DC 20208

1 Dr. Thomas G. Sticht
Basic Skills Program
National Institute of Education
1200 19th Street NW
Washington, DC 20208

Civil Govt.

1 Dr. Vern W. Urry
Personnel R & D Center
U.S. Civil Service Commission
1900 E. Street NW
Washington, DC 20415

1 Dr. Joseph L. Young, Director
Memory & Cognitive Processes
National Science Foundation
Washington, DC 20550

Non Govt.

1 Dr. Earl A. Alluisi
HQ, AFHRL (AFSC)
Brooks AFB, TX 78235

1 Dr. Erling B. Anderson
University of Copenhagen
Studiestraedt
Copenhagen
DENMARK

1 1 psychological research unit
Dept. of Defense (Army Office)
Campbell Park Offices
Canberra ACT 2600
AUSTRALIA

1 Dr. Alan Baddeley
Medical Research Council
  Applied Psychology Unit
15 Chaucer Road
Cambridge CB2 2EF
ENGLAND

1 Dr. Isaac Bejar
Educational Testing Service
Princeton, NJ 08450

Non Govt.

1 Dr. Warner Birice
Streitkraefteamt
Rosenberg 5300
Bonn, West Germany d-5300

1 Dr. R. Darrel Bock
Department of Education
University of Chicago
Chicago, IL 60637

1 Dr. Nicholas A. Bond
Dept. of Psychology
Sacramento State College
600 Jay Street
Sacramento, CA 95819

1 Dr. David G. Bowers
Institute for Social Research
University of Michigan
Ann Arbor, MI 48106

1 Dr. Robert Brennan
American College Testing Programs
P.O. Box 168
Iowa City, IA 52240

1 Dr. C. Victor Bunderson
WICAT INC.
University Plaza, Suite 10
1160 S. State St.
Orem, UT 84057

1 Dr. John B. Carroll
Psychometric Lab
University of N. Carolina
Davie Hall 013A
Chapel Hill, NC 27514

1 Charles Myers Library
Livingstone House
Livingstone Road
Stratford
London E15 2LJ
ENGLAND

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

Non Govt.

1 Dr. Norman Cliff
Department of Psychology
Univ. of S. California
University Park
Los Angeles, CA 90007

1 Dr. William Coffman
Iowa Testing Programs
University of Iowa
Iowa City, IA 52242

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

1 Dr. Meredith Crawford
Department of Engineering Adm.
George Washington University
Suite 805
2101 L Street NW
Washington, DC 20037

1 Dr. Hans Cronbag
Education Research Center
University of Leyden
Boerhaavelaan 2
Leyden
The NETHERLANDS

1 Major I. N. Evonic
Canadian Forces Pers. Applied
Research
1107 Avenue Road
Toronto, Ontario
CANADA

1 Dr. Leonard Feldt
Lindquist Center for Measurement
University of Iowa
Iowa City, IA 52242

1 Dr. Richard L. Ferguson
The American College Testing
Program
P.O. Box 168
Iowa City, IA 52240

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

Non Govt.

1 Dr. Gerhardt Fischer
Liebigasse 5
Vienna 1010
AUSTRIA

1 Dr. Donald Fitzgerald
University of New England
Armidale, New South Wales 2351
AUSTRALIA

1 Dr. Edwin A. Fleishman
Advanced Research Resources Org.
Suite 900
4330 East West Highway
Washington, DC 20014

1 Dr. John R. Frederiksen
Bolt Beranek & Newman
50 Moulton Street
Cambridge, MA 02138

1 Dr. Robert Glaser
LRDC
University of Pittsburgh
3939 O'Hara Street
Pittsburgh, PA 15213

1 Dr. Ross Greene
CTB/McGraw Hill
Del Monte Research Park
Monterey, CA 93940

1 Dr. Alan Gross
Center for Advanced Study
in Education
City University of New York
New York, NY 10036

1 Dr. Ron Hambleton
School of Education
University of Massachusetts
Amherst, MA 01002

1 Dr. Chester Harris
School of Education
University of California
Santa Barbara, CA 93106

1 Dr. Lloyd Humphreys
Department of Psychology
University of Illinois
Champaign, IL 61820

Non Govt.

1 Library
HumRRO/Western Division
27857 Berwick Drive
Carmel, CA 93921

1 Dr. Steven Hunka
Department of Education
University of Alberta
Edmonton, Alberta
CANADA

1 Dr. Earl Hunt
Department of Psychology
University of Washington
Seattle, WA 98105

1 Dr. Huynh Huynh
Department of Education
University of South Carolina
Columbia, SC 29208

1 Dr. Carl J. Jensema
Gallaudet College
Kendall Green
Washington, DC 20002

1 Dr. Arnold F. Kanarick
Honeywell, Inc.
2600 Ridgeway Parkway
Minneapolis, MN 55413

1 Dr. John A. Keats
University of Newcastle
Newcastle, New South Wales
AUSTRALIA

1 Mr. Marlin Kroger
1117 Via Goleta
Palos Verdes Estates, CA 90274

1 LCOL. C. R. J. Lafleur
Personnel Applied Research
National Defense HQS
101 Colonel by Drive
Ottawa, CANADA K1A OK2

1 Dr. Michael Levine
Department of Psychology
University of Illinois
Champaign, IL 61820

Non Govt.

1 Dr. Robert Linn
College of Education
University of Illinois
Urbana, IL 61801

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

1 Dr. Gary Marco
Educational Testing Service
Princeton, NJ 08540

1 Dr. Scott Maxwell
Department of Psychology
University of Houston
Houston, TX 77025

1 Dr. Sam Mayo
Loyola University of Chicago
Chicago, IL 60601

1 Dr. Allen Munro
USC
Behavioral Technology Labs
3717 South Hope Street
Los Angeles, CA 90007

1 Dr. Melvin R. Novick
Iowa Testing Programs
University of Iowa
Iowa City, IA 52242

1 Dr. Jesse Orlansky
Institute for Defense Analysis
400 Army Navy Drive
Arlington, VA 22202

1 Dr. James A. Paulson
Portland State University
P.O. Box 751
Portland, OR 97207

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

1 Dr. Steven M. Pine
4950 Douglas Avenue
Golden Valley, MN 55416

Non Govt.

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

1 Min. Ret. M. Rauch
P II 4
BUNDESMINISTERIUM DER VERTEIDIGUNG
Postfach 161
53 Bonn 1, GERMANY

1 Dr. Peter B. Read
Social Science Research Council
605 Third Avenue
New York, NY 10016

1 Dr. Mark D. Reckase
Educational Psychology Dept.
University of Missouri-Columbia
4 Hill Hall
Columbia, MO 65211

1 Dr. Andrew M. Rose
American Institutes for Research
1055 Thomas Jefferson St. NW
Washington, DC 20007

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

1 Dr. Ernst Z. Rothkopf
Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974

1 Dr. Donald Rubin
Educational Testing Service
Princeton, NJ 08450

1 Dr. Larry Rudner
Gallaudet College
Kendall Greeen
Washington, DC 20002

1 Dr. J. Ryan
Department of Education
University of South Carolina
Columbia, SC 29208

Non Govt.

1 Prof. Fumiko Samejima
Dept. of Psychology
University of Tennessee
Knoxville, TN 37916

1 Dr. Robert J. Seidel
Instructional Technology Group
HUMRRO
300 N. Washington Street
Alexandria, VA 22314

1 Dr. Kazao Shigemasu
University of Tohoku
Department of Ed. Psych.
Kawauchi, Sendai 982
JAPAN

1 Dr. Edwin Shirkey
Department of Psychology
Florida Technological Univ.
Orlando, FL 32816

1 Dr. Richard Snow
School of Education
Stanford University
Stanford, CA 94305

1 Dr. Robert Sternberg
Dept. of Psychology
Yale University
Box 11A, Yale Station
New Haven, CT 06520

1 Dr. Albert Stevens
Bolt Beranek & Newman, Inc.
50 Moulton Street
Cambridge, MA 02138

1 Dr. Patrick Suppes
Institue for Mathematical Studies
in the Social Sciences
Stanford University
Stanford, CA 94305

1 Dr. Hariharan Swaminathan
Laboratory of Psychometric &
Evaluation Research
School of Education
University of Massachusetts
Amherst, MA 01003

Non Govt.

1 Dr. Brad Sympson
Elliott Hall
University of Minnesota
75 E. River Road
Minneapolis, MN 55455

1 Dr. Kikumi Tatsuoka
Computer Based Education
Research Laboratory
252 Engineering Research Lab.
University of Illinois
Urbana, IL 61801

1 Dr. David Thissen
Department of Psychology
University of Kansas
Lawrence, KS 66044

1 Dr. Robert Tsutakawa
Dept. of Statistics
University of Missouri
Columbia, MO 65211

1 Dr. J. Uhlaner
Perceptronics, Inc.
6271 Variel Avenue
Woodland Hills, CA 91364

1 Dr. Howard Wainer
Bureau of Social Science Research
1990 M. Street, NW
Washington, DC 20036

1 Dr. Thomas Wallsten
Psychometric Laboratory
Davie Hall 013A
University of North Carolina
Chapel Hill, NC 27514

1 Dr. John Wannous
Department of Management
Michigan University
East Lansing, MI 48824

1 Dr. David J. Weiss
N660 Elliott Hall
University of Minnesota
75 E. River Road
Minneapolis, MN 55455

Non Govt.

1 Dr. Susan Whitely
Psychology Department
University of Kansas
Lawrence, KA 66044

1 Dr. Wolfgang Wildgrube
Streitkraefteamt
Rosenberg 5300
Bonn, WEST GERMANY D-5300

1 Dr. Robert Woud
School Examination Department
University of London
66-72 Gower Street
London WCIE 6EE
ENGLAND

1 Dr. Karl Zinn
Center for Research on
Learning & Teaching
University of Michigan
Ann Arbor, MI 48104