DOCUMENT RESUME

ED 177 196 TM 009 669

AUTHOR	Koch, William R.; Reckase, Mark D.
TITLE	Problems in Application of Latent Trait Models to Tailored Testing.
INSTITUTION	Missouri Univ., Columbia.
SPONS AGENCY	Office of Naval Research, Washington, D.C. Personnel and Training Branch.
PUB DATE	Apr 79
CONTRACT	N00014-77-C0097
NOTE	23p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, California, April 9-11, 1979)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Achievement Tests; Educational Testing; Higher Education; Item Analysis; *Item Banks; *Mathematical Models; Multiple Choice Tests; Psychometrics; *Test Construction; Test Reliability; Test Validity; Verbal Tests; Vocabulary Skills
IDENTIFIERS	*Computer Assisted Test Construction; Latent Trait Models; One Parameter Model; Tailored Testing; Three Parameter Model

ABSTRACT

Tailored testing procedures for achievement testing were applied in a situation that failed to meet some of the specifications generally considered to be necessary for tailored testing. Discrepancies from the appropriate conditions included the use of small samples for calibrating items, and the use of an item pool that was not designed to be homogeneous in content. The item pool contained 180 items concerning educational measurement that were calibrated separately by a one-parameter logistic model and by a three-parameter logistic model. The 110 undergraduate students were each tested at two sessions a week apart with both one-parameter and three-parameter tailored tests at each session. All tests were administered on a computer terminal. The results were studied for several characteristics including: goodness of fit of the observed responses to those predicted by each model: the information function of each test compared to a fifty-item traditional paper-and-pencil test; the reliabilities of the tailored tests; and the content validities of the tailored tests. A unidimensional tailored test of vocabulary was also administered, with satisfactory results. The achievement tests generally produced unsatisfactory results, presumably because of the discrepancies from appropriate conditions. (CTM)

Problems in Application of Latent Trait Models to Tailored Testing

by

William R. Koch and Mark D. Reckase
University of Missouri-Columbia

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.

Tailored testing procedures (Lord, 1970; Weiss, 1974) have been developed to alleviate these and other problems with conventional tests, but we will see that, in so doing, a whole new host of problems may be introduced. The purpose of the present paper is to describe some of these 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 primary characteristics.

One major 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 procedures attempt to "tailor make" the test for each individual. This is accomplished by the selection of items for administration that approximately match item difficulty parameters to an examinee's estimated ability level after each response to an item, resulting in efficient measurement that facilitates the control of test errors.

However, in order to implement tailored testing it is usually necessary to utilize computer capabilities for several steps in the procedure.

Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, April, 1979. This research was supported by Contract Number N00014-77-C0097 from the Personnel and Training Research Programs of the Office of Naval Research.

Tailored testing itself is often based on latent trait or item characteristic curve (ICC) theory (Lord, 1952; Lord and Novick, 1968) which involve relatively sophisticated mathematical models. In addition, the procedures require a precalibrated pool of items to be available for selecting the test items to be administered. This 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 in an actual test setting 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 paper, 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 in 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

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

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 certain response to any given item on a test is unaffected by any previous response.

The unidimensionality assumption has particular relevance when consideringing 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 72 item 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 based on the information function (Birnbaum, 1968), and maximum likelihood ability estimation was used.

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 test 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 than the 1PL model.

In regard to reliability, the 3PL procedure 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 cannot be too highly emphasized that the 3PL procedure, in conjunction with maximum likelihood ability estimation, failed to converge at 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 does not occur in conjunction with the 1PL model.

Tailored Achievement Testing

One interesting application of ICC theory was reported by Brown and Weiss (1977) in which a tailored testing procedure was used for an achievement test with multiple 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. However, this application to multidimensional achievement measurement 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 investigate another crucial aspect of achievement testing, namely content validity. The current study provided an opportunity to examine both the robustness of the ICC model and the content validity of tailored achievement testing.

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.

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 indexes, and in such a manner that the average test difficulties were close to .75. Being achievement t-sts, a table of specifications was used to construct the tests to match course objectives. KR-20 reliabilities for the exams were consistently found to be in the range from +.60 to +.85.

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 is to link all the separate item parameter calibrations into one final set of item parameters such that parameter estimates obtained from different samples are 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 as well.

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. But it is important to note that the parameter estimates are equivalent within a linear transformation. This means that the very desirable attribute of latent trait or ICC models referred to as invariance of item parameters (Lord and Novick, 1968) is still maintained.

For space reasons, the present paper will only briefly describe the procedures used to link the separate item calibrations together into one large pool of 180 items for tailored testing. (Reckase, 1979, provides a thorough discussion of item linking techniques.) In the current study one of the tests was arbitrarily designated as the calibration base for linking. Then the 3PL item discrimination estimates and the 1PL item difficulty estimates were linked from all the separate test calibrations onto the same scales as their corresponding item parameters in the calibration base. The linear transformation incorporated the use of multiplicative constants in the case of the 3PL linking and additive constants in the case of the 1PL linking. The 3PL item difficulty parameters were linked by means of simple linear regression. The 3PL item guessing parameters required no transformation since they were already on the same 0 to 1 scale, but they were combined using a weighted average procedure.

Table 1 presents the means, standard deviations, and ranges of the item parameter estimates resulting from the calibration and linking procedures described above. Both the 1PL and 3PL item pools contained exactly the same items. The correlation between their respective item difficulty parameters was .91. The distributions of the difficulty parameters were markedly peaked rather than taking on a uniform distribution which would have been preferred, based on previous research.

Insert Table 1 about here

Tailored Testing Procedures

The three required components of the tailored testing procedure included (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 the 3PL procedures, items were selected for administration which maximized 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

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

where $I(\theta_j, u_{ij})$ is the information of Item i at ability level θ for Person j, given item response u_{ij}, with 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

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

where $I(\theta_j, u_{ij})$ 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:

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

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 for the initial ability estimate. If the examinee answered the first item correctly, the new ability estimate was placed at a fixed stepsize (.693) away in a positive direction (i.e. a more difficult item). An incorrect response resulted in an ability estimate that was -.693 away. 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 estimate. When at least one correct and one incorrect response were obtained, the ability level of the examinee was 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 procedures cycled through this process until one of two stopping rules was reached: either no item remained in the item pool with an information value greater than a specified amount, or a maximum of 20 items had been administered.

Design

The study employed a counterbalanced designed 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 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. A total of 110 students took part.

Analyses

The primary research issues in the 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, and the correlation of the ability estimates with the conventional course exam.

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. 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 differenced were computed using the formula

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

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.

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 to the information provided by the traditional 50 item paper-and-pencil course exam. Again, the plot constructed for the relative efficiency comparison was based on a selected sample of cases across the whole range of tailored testing ability estimates.

The content validity analyses were conducted to determine the degree to which both the item pools and the tailored tests accurately represented the measurement of the course objectives that had been specified. Since a table of specifications was used to construct the traditional course exam, a particular weighting of test items to content areas was assured. The issue was whether or not the item pools and tailored tests reflected the same weightings. A set of Chi Square analyses were performed to determine the fit between desired and observed content distributions in this respect.

Other analyses included descriptive statistics for the two types of tailored tests, including average test length, average test difficulty, number of items actually used from the item pools, etc. In addition, several correlations were computed, such as ability estimate intercorrelations across models and correlations of tailored test scores with regular exam scores. Finally, a principal components analysis of the traditional course exam was run to determine its structure. The purpose was to determine if the achievement test was truly multidimensional.

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 1.

Insert Table 2 about here

The horizontal broken line indicates the information of the traditional 50 item course achievement test 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. Even so, a subjective visual comparison of the two is possible.

Insert Figure 1 about here

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 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 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 ($p<.01$) 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 a frequently encountered problem when using 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.)

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.)

Insert Table 3 about here

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. 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 dificult 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 2 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 happen to answer the first 3 or 4 tailored tests items correctly and "top out" 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.

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 indicated that the proportion of items in common had a negligible effect on the test reliability in previous research.

Other Correlation Analyses

In Table 4 are listed the correlations computed between the tailored test ability estimates and scores on the paper-and-pencil course exams. In general, the correlations were relatively low. This was also true for the tailored test correlations with Exam 1, even though the tests covered the same content areas.

Insert Table 4 about here

Descriptive Statistics

Insert Table 5 about here

Table 5 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 tests, the values for the mean number of items administered indicate that most of the tests went the full distance. This result implied that ample numbers of items were available in the item pool which had sufficient information for most of the examinees. The mean test difficulty values reflected the overall low difficulty of the items for the majority of the students, since the mean proprotion 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.

Content Validity

Insert Table 6 about here

As can be seen in Table 6, 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 just 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.

DISUCSSION

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 predicted 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, which 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 test 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 ability 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" 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, where long strings of incorrect responses were obtained. In such a case no reasonable maximum likelihood ability estimate could be calculated since the likelihood function approached a uniform distribution with the mode 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 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. 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 should not necessarily be expected to reflect the same weightings. For the 1PL procedure this result was somewhat of a surprise, if the assumption is made that ability is normally distributed. In such a case, 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. 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.

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; 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.

FIGURE 1

RELATIVE EFFICIENTY

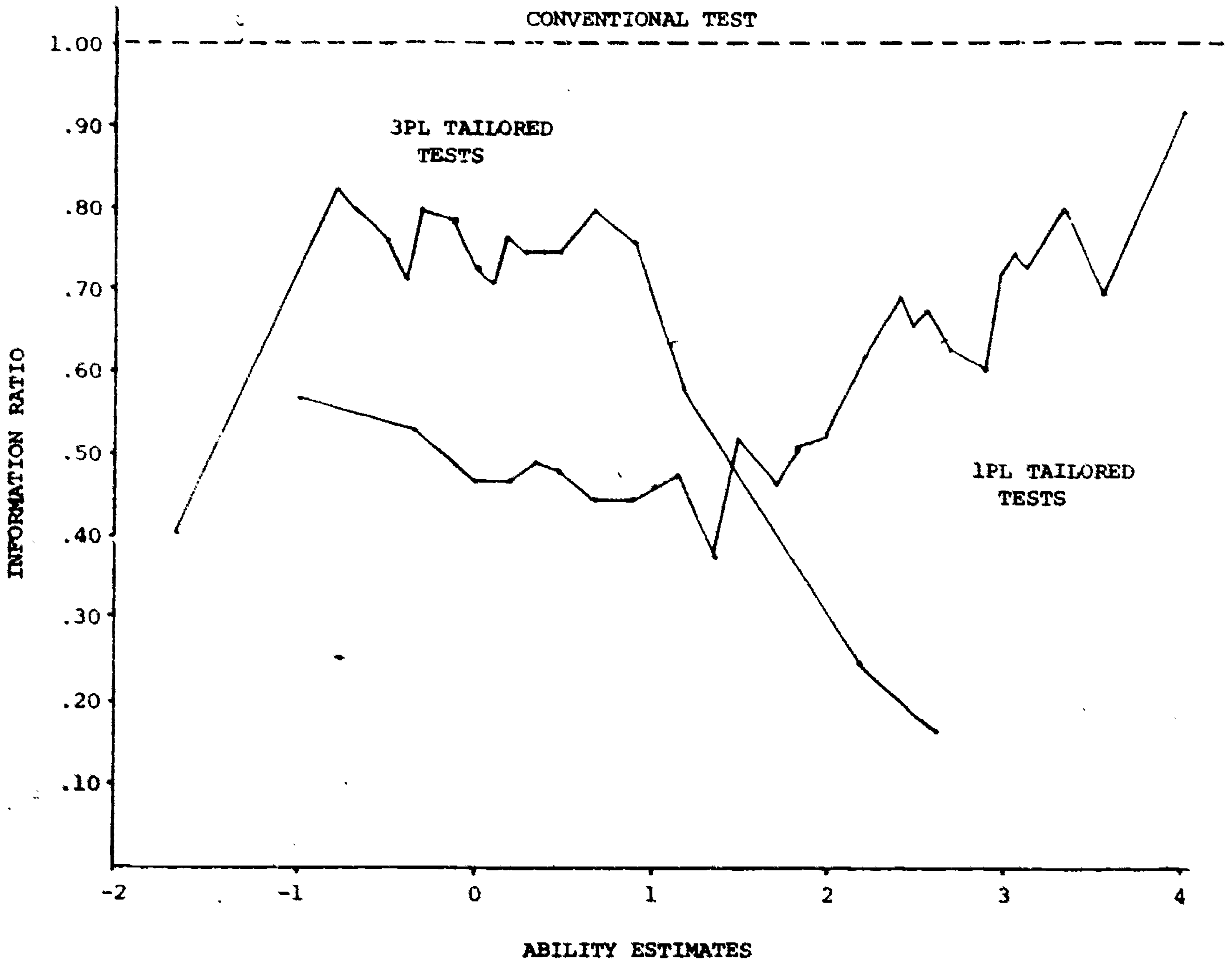

Table 1

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 artificial lower limit on the 3PL difficulty parameters.

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
$\bar{X}$	.2046	.1649
$S_{\bar{x}}$	.0426	.0701

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

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 non-convergence cases)

Table 4

Correlations of Ability Estimates
With Traditional Course Exams[a]

Variables	1PL 1	1PL 2	3PL 1	3PL 2
Exam 1	.30	.41	.42	.09
Exam 2	.35	.28	.17	.20
Exam 3	.31	.22	.27	.20
Total Score	.57	.48	.41	.23

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

Table 5

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)

Table 6

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
Behavior 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 Specs.	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, $\chi^2 = 133.448$

REFERENCES

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. 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 of 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.