# USEFUL TECHNIQUES FOR APPLYING LATENT TRAIT MENTAL-TEST THEORY

by

Carl Jensema

TECHNICAL REPORT NO. 202

May 9, 1973

PSYCHOLOGY & EDUCATION SERIES



INSTITUTE FOR MATHEMATICAL STUDIES IN THE SOCIAL SCIENCES

STANFORD UNIVERSITY

STANFORD, CALIFORNIA

Abstract

A simple method for estimating parameters for the Birnbaum three-parameter logistic mental-test model was outlined. The accuracy of the method was investigated with Monte Carlo data, and advantages and disadvantages are pointed out and discussed. Data from six vocabulary tests were then used to demonstrate the usefulness of the method in prescreening items for inclusion in potential tailored testing-item banks.

# USEFUL TECHNIQUES FOR APPLYING LATENT TRAIT MENTAL-TEST THEORY[1]

Carl Jensema[2]

## 1. Introduction

A new type of mental-test theory has evolved over the past thirty years. Beginning with the pioneering work of Lord (1952, 1953, 1964, 1968, 1970a, 1970b, 1971), Birnbaum (1957a, 1957b, 1957c, 1968), and others, an impressive body of statistical theory now exists which is known as the latent trait mental-test theory.

The concept of Item Characteristic Curves (ICC) is basic to latent trait mental-test theory. In brief, when the probability of passing an item is plotted against a scale of ability, the resulting plot resembles a normal-ogive. This curve is referred to as the ICC of the item.

Lord (1970a) has proposed a simple technique for approximating ICCs without knowledge of their mathematical form. In this method, the observed score is taken as the total number of items answered correctly, excluding the item to be plotted. Distributions of true scores are then estimated from the observed-score distribution coupled with suitable assumptions. The proportion of examinees responding correctly to the item is plotted against the true score values. This gives an estimate of the regression curve of item score on true score. Since the true score on a test may be viewed as a monotonic transform of the latent trait measured by the test, the estimated regression curve can be transformed into an estimate of the ICC. Lord found close agreement between

plots obtained in this manner and theoretical curves based on Birnbaum's three-parameter logistic mental test model (Birn-3 model). As will be seen later, this agreement has important implications for practical application of the Birn-3 model.

In mathematical terms, the Birn-3 model assumes that the probability of an examinee of ability $\theta$ responding correctly to item i is

$$(1) \qquad P_i'(\theta) = c_i + \frac{1 - c_i}{1 - \exp\left[-Da_i(\theta - b_i)\right]}$$

where D is a constant taken as 1.7 and the parameters $a_i$, $b_i$, and $c_i$ are related to discriminating power, difficulty, and probability of guessing correctly for item i. The probability of a wrong answer to item i by an examinee of ability $\theta$ is simply

$$(2) \qquad Q_i'(\theta) = 1 - P_i'(\theta)$$

$$= \frac{1 - c_i}{\left[1 + \exp\left[DA_i(\theta - b_i)\right]\right]} .$$

Detailed discussions of the mathematical basis of the Birn-3 and other models may be found in Birnbaum (1968) and Jensema (1972).

If the values of $a_i$, $b_i$, and $c_i$ are known for an item we, in effect, known its ICC. If there are a large number of these items (say 100), we may create an item bank and apply tailored testing techniques, which permit consideration of the characteristics of individual items. In conventional testing, most items are too easy or too hard for the ability of a given examinee. In tailored testing, only those items appropriate for a given level of ability are administered. The interested reader

may consult Lord (1970b, 1971), Owen (1969), Urry (1971b), and Jensema (1972) for discussions on these techniques.

At the present time, the Birn-3 is the most practical latent trait mental-test model available for tailored testing. The biggest stumbling block has been the difficulty of estimating parameter values for the items (Lord, 1968; Jensema, 1972). Not only is the standard technique of using maximum likelihood methods mathematically complex and hard to program on a computer, but it also consumes formidable amounts of computer time. The expenditure of hundreds or even thousands of dollars' worth of computer time to routinely answer the simple questions of which items in an item pool should be included in a tailored testing-item bank is difficult to justify.

In this paper I shall discuss some simple techniques for the economical estimation of Birn-3 item parameters. I shall apply these methods and show that conventional tests are a potentially good source of items for a tailored testing item bank.

## 2. Method

The first difficulty in estimating item parameter values for the Birn-3 is in determining the guessing parameter value ($c_i$). From a multiple-choice item with five possible responses, we would expect a $c_i$ value of .20 for an examinee who guessed randomly. However, because of distractor responses, "don't guess" instructions to the examinee, and a variety of other reasons, the value of the guessing parameter is often lower than what would be mathematically expected.

The $c_i$ value can be obtained through maximum likelihood techniques in the same manner as the other parameter values. Lord (1968) tried this

and found that maximum likelihood estimation of $c_i$ values did not work well. Convergence was slow and absurd values were sometimes obtained. He recommended obtaining "eyeball" estimates of $c_i$ from the asymptotic value of the lower tail of the approximated ICC. Actually, it is possible to obtain the proportion of examinees passing an item at each of the lower item-excluded subtest scores without plotting an approximate ICC. By examining these proportions, one can get a fairly good approximation of the $c_i$ value of that item.

Obtaining approximations of $a_i$, the discriminatory power, and $b_i$, the difficulty, for an item is a bit more complex. First, consider that $P_i'(\theta)$, the probability of a correct response to item i with respect to ability $\theta$, is the sum of the probability of knowing the correct response and the probability of not knowing the correct response but guessing correctly. Then we may express $P_i'(\theta)$ as

$$P_i'(\theta) = P_i(\theta) + c_i[1 - P_i(\theta)] \quad . \tag{3}$$

We may manipulate (3) to obtain the probability of knowing the correct response as

$$P_i(\theta) = \frac{P_i'(\theta) - c_i}{1 - c_i} \quad . \tag{4}$$

In the general model (Birnbaum, 1968), $P_i(\theta)$ is expressed as

$$P_i(\theta) = \frac{1}{\sqrt{2\pi}} \int_{\gamma_i}^{\infty} \exp \frac{-t^2}{2} \, dt \tag{5}$$

where $\gamma_i$ is the correct-incorrect cut point for item i on an unobservable

response continuum. We may obtain a close approximation of (5) by using a logistic function:

$$P_i(\theta) = \frac{1}{1 + \exp\,[D\gamma_i]} \tag{6}$$

where D = 1.7, a constant. Manipulation of (6) allows us to obtain $\gamma_i$ as

$$\gamma_i = \frac{1}{D}\,\ell n\,\frac{1 - P_i(\theta)}{P_i(\theta)}\ . \tag{7}$$

The height of the ordinate of a normal curve at point $\gamma_i$ is

$$\phi(\gamma_i) = \frac{1}{\sqrt{2\pi}}\exp\left[\frac{-\gamma_i^2}{2}\right]\ . \tag{8}$$

Let $\rho_{I\theta}$ represent the correlation between the unobservable response continuum and underlying ability. Then the covariance between binarily scored item i and underlying ability $\theta$ is

$$\sigma_{i\theta} = \rho_{I\theta}\,\phi(\gamma_i)\ . \tag{9}$$

From (4) the variance of a binary item is

$$\sigma_i^2 = P_i(\theta)\,[1 - P_i(\theta)]\ . \tag{10}$$

By dividing (9) by the square root of (10) we obtain the point biserial correlation between the binarily scored response to item i and ability as

$$\rho_{i\theta} = \frac{\sigma_{i\theta}}{\sigma_i} = \frac{\rho_{I\theta}\phi(\gamma_i)}{\sqrt{P_i(\theta)[1-P_i(\theta)]}}\ . \tag{11}$$

Note that this is for the free-response situation where the probability of guessing correctly is zero.

To take the effects of guessing into consideration we express (9) and (10) as

$$\sigma'_{i\theta} = (1-c_i)\ \rho_{I\theta}\ \phi(\gamma_i) \tag{12}$$

and

$$\sigma_i^{2'} = P'_i(\theta)\ [1 - P'_i(\theta)]\ . \tag{13}$$

Then (11) becomes

$$\rho'_{i\theta} = \frac{\sigma'_{i\theta}}{\sigma'_i} = \frac{\rho_{I\theta}\ \phi(\gamma_i)(1 - c_i)}{\sqrt{P'_i(\theta)[1 - P'_i(\theta)]}}\ . \tag{14}$$

By manipulating (14) we may express $\rho_{I\theta}$ as

$$\rho_{I\theta} = \frac{\rho'_{i\theta}\ \sqrt{P'_i(\theta)[1 - P'_i(\theta)]}}{\phi(\gamma_i)(1 - c_i)}\ . \tag{15}$$

The model defines the relationship between $\rho_{I\theta}$ and $a_i$ as

$$\rho_{I\theta} = \frac{a_i}{\sqrt{1 + a_i^2}}\ , \tag{16}$$

which may also be written as

$$a_i = \frac{\rho_{I\theta}}{\sqrt{1 - \rho_{I\theta}^2}}\ . \tag{17}$$

Further, the relationship between $\gamma_i$, $\rho_{I\theta}$, and $b_i$ is given by the model as

(18) $$\gamma_i = \rho_{I\theta} b_i$$

or, alternatively, as

(19) $$b_i = \frac{\gamma_i}{\rho_{I\theta}} .$$

The practical application of the equations in the preceding paragraphs is straightforward and simple. An estimate of $c_i$ may be obtained for item i by examining the proportion of examinees passing item i at each item-excluded subtest score. Consider $P_i'(\theta)$ as the proportion of the population passing item i. The point biserial correlation, $\rho_{i\theta}$, between the binarily scored responses to item i and underlying ability may be calculated by assuming that $\theta$ is represented by the item-excluded test score. Then, by utilizing (4), (7), (8), (15), (17), and (19), we may estimate the values of $a_i$ and $b_i$.

To demonstrate just how simple this is, consider C, P, and R as estimates of $c_i$, $P_i'(\theta)$, and $\rho_{i\theta}$, respectively. The following FORTRAN commands will yield estimates of $a_i$ and $b_i$:

```
PI = 3.1415926535897793
PP = (P - C)/(1. - C)
CUT = ALOG ((1. - PP)/PP)/1.7
HEIGHT = EXP (CUT**2/(-2))/SQRT(2.*PI)
R = (R*SQRT(P*(1. - P)))/(HEIGHT*(1. - C))
A = R/SQRT(1. - R**2)
B = CUT/R
```

The major weakness in the method is that underlying ability $\theta$ is clearly not the same as the item-excluded test score $X_{(i)}$. The point

biserial correlation between a binary item and the item-excluded test score is (see Lord & Novick, 1968, Ch. 15.5)

$$r_{i\chi_{(i)}} = \frac{\bar{\chi}^{+}_{(i)} - \bar{\chi}^{-}_{(i)}}{\sigma_{\chi_{(i)}}} \sqrt{P_{\chi_{(i)}}(1 - P_{\chi_{(i)}})} \tag{20}$$

where $P_{\chi_{(i)}}$ is the probability of passing item i at subtest score $\chi_{(i)}$, $\bar{\chi}^{+}_{(i)}$ is the mean score for those passing item i, and $\bar{\chi}^{-}_{(i)}$ is the mean score for those failing item i. The point biserial correlation between the binary item and the underlying ability is

$$\rho_{i\theta} = \frac{\bar{\theta}^{+} - \bar{\theta}^{-}}{\sigma_{\theta}} \sqrt{P_i(\theta)(1 - P_i(\theta))} \, . \tag{21}$$

If $\chi_{(i)}$ had linear regression on $\theta$,

$$\mathcal{E}(\chi_{(i)}|\theta) = \alpha\theta + \beta \tag{22}$$

and then

$$\frac{\bar{\chi}^{+}_{(i)} - \bar{\chi}^{-}_{(i)}}{\sigma_{\chi_{(i)}}} \rightarrow \frac{\alpha\bar{\theta}^{+} + \beta - \alpha\bar{\theta}^{-} - \beta}{\sqrt{\alpha\sigma_{\theta}^2 + \sigma_e^2}} \tag{23}$$

where $\sigma_e^2$ is the error variance. If $\sigma_e^2$ is small $r_{i\chi_{(i)}}$ should approximate $\rho_{i\theta}$, since

$$\frac{\bar{\chi}^{+}_{(i)} - \bar{\chi}^{-}_{(i)}}{\sigma_{\chi_{(i)}}} \cong \frac{\bar{\theta}^{+} - \bar{\theta}^{-}}{\sqrt{\sigma_{\theta}^2 + \sigma_e^2}} \, . \tag{24}$$

However, the regression of $\chi_{(i)}$ on $\theta$ is almost certainly nonlinear, and $r_{i\chi_{(i)}}$ may be regarded only as a rough first-order approximation of $\rho_{i\theta}$.

It is to be expected that for extreme values $\mathcal{E}(\chi_{(i)}|\theta) \neq \alpha\theta + \beta$ even approximately. The method described in this paper uses $r_{i\chi_{(i)}}$ to approximate $\theta$ because of its simplicity and because, as will be seen, it can lead to fairly good results in practice.

The method is a modification of a more cumbersome graphic technique proposed by Urry (1971a). Jensema (1972) utilized Urry's graphs to obtain initial $a_i$ and $b_i$ estimates for maximum likelihood parameter estimation. The correlations between the graphic estimates and the maximum likelihood estimates were .89 and .97 for $a_i$ and $b_i$, respectively.

It was decided to test the present parameter estimation method with Monte-Carlo techniques. A computer program described by Urry (1970) was used to generate random data for 500 "examinees" on 100 "items" with known parameter values. The items had 25 difficulty levels (-2.4, -2.2, ..., 0, 2.2, 2.4), and each difficulty level had four items with discriminatory values set to .5, 1.0, 1.5, and 2.0, respectively. The guessing probability was set to .20 for all items. The ability of the Monte-Carlo examinees created by the program was normally distributed to assure that the data fit the model in this respect. Estimates of the discrimination and difficulty parameters for each of the 100 Monte-Carlo items were calculated by the method outlined. The parameter estimates were then correlated with the known parameter values.

The next step involved the creation of a potential item bank from real data. The most obvious place to obtain items for a tailored testing item bank based on the Birn-3 is in conventional multiple-choice tests. Unfortunately, it has been clearly demonstrated (Lord, 1968; Jensema, 1972) that not all conventional test items are good tailored testing

items. Further, most conventional tests do not have enough items for a good item bank, even if all items are good for tailored testing.

One way of handling this is to obtain data from similar tests, evaluate the parameters of the items in each separate test, discard items unsuitable for tailored testing, and combine the remaining items into a single item pool. This pool could be administered to a group of examinees, and the parameters could be reestimated to obtain a final item bank.

Data on six vocabulary tests were obtained through the California Test Bureau. The tests were selected from the Comprehensive Tests of Basic Skills (CTBS), levels 2, 3, and 4, and the California Achievement Tests (CAT), levels 3, 4, and 5. Note that it is possible to pool items from different levels because of item parameter invariance. These particular tests were chosen partly because the simplicity of the items would make them easy to administer by computer if a final usable tailored testing item bank resulted. The various levels of the tests were originally designed for grades 4 through 12. The data were from separate groups of 5,400 subjects for each test. Each of the six tests had 40 items. However, because of a computer programming error, data for the last item in each test were lost, and thus 39 rather than 40 items were evaluated for each test.

The values of $a_i$ and $b_i$ were calculated by the method described earlier. These estimates were examined, selection criteria were established, and items not meeting these criteria were eliminated. The remaining items were combined into a potential item bank for future research.

## 3. Results

The results of the parameter estimation with 500 Monte-Carlo "examinees" are given in Table 1. It can readily be observed from the table that parameter estimation tended to be most accurate for items with low discriminatory power and with a difficulty level roughly in the -1.0 to 1.0 range.

For item 25, the estimates in Table 1 are quite unreasonable. A discriminatory power of zero on this item means that the estimated item characteristic curve would simply be a horizontal line. Examination of the Monte-Carlo data indicated that item 25 was so difficult and had so little discriminatory power that the random data generated for it had, as might be expected, a zero correlation with the subtest of other items.

For 9 Monte-Carlo items, the discriminatory power could not be estimated. Examination of Equation (17) indicates that this is caused by $\rho_{I\theta}$ values which exceed 1.0. Since correlations can never exceed 1.0, such $\rho_{I\theta}$ values are obviously erroneous. They tend to occur most frequently when the correlation between an item and an item-excluded subtest is low, and when the proportion of the population passing the item is either very high or very low. Inability to estimate discrimination parameter values under some circumstances is characteristic of the method which limits its practical application.

The correlation between true and estimated discrimination power was calculated (excluding item 25 and those items for which an estimate could not be calculated) and was found to be .59. The correlation between true and estimated difficulty (excluding item 25) was .97. When only those items with true difficulty values between -1.0 and +1.0 were

TABLE 1

Known and Estimated Parameter Values for Monte-Carlo Data

(c = .20 for all items)

| Item number | a | b | $\hat{a}$ | $\hat{b}$ | Item number | a | b | $\hat{a}$ | $\hat{b}$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | .5 | -2.4 | .35 | - 3.06 | 26 | 1.0 | -2.4 | 1.21 | -2.14 |
| 2 | .5 | -2.2 | .53 | - 1.81 | 27 | 1.0 | -2.2 | 1.11 | -1.76 |
| 3 | .5 | -2.0 | .45 | - 2.34 | 28 | 1.0 | -2.0 | 1.49 | -1.43 |
| 4 | .5 | -1.8 | .53 | - 1.46 | 29 | 1.0 | -1.8 | 1.05 | -1.70 |
| 5 | .5 | -1.6 | .66 | - 1.11 | 30 | 1.0 | -1.6 | .68 | -1.72 |
| 6 | .5 | -1.4 | .40 | - 1.58 | 31 | 1.0 | -1.4 | 1.04 | -1.28 |
| 7 | .5 | -1.2 | .61 | - 1.02 | 32 | 1.0 | -1.2 | .87 | -1.27 |
| 8 | .5 | -1.0 | .61 | - .73 | 33 | 1.0 | -1.0 | 1.14 | - .76 |
| 9 | .5 | - .8 | .59 | - .70 | 34 | 1.0 | - .8 | 1.08 | - .69 |
| 10 | .5 | - .6 | .51 | - .55 | 35 | 1.0 | - .6 | .93 | - .50 |
| 11 | .5 | - .4 | .34 | - .40 | 36 | 1.0 | - .4 | 1.14 | - .26 |
| 12 | .5 | - .2 | .48 | - .22 | 37 | 1.0 | - .2 | .96 | - .14 |
| 13 | .5 | 0.0 | .64 | .16 | 38 | 1.0 | 0.0 | .82 | - .11 |
| 14 | .5 | .2 | .52 | .46 | 39 | 1.0 | .2 | 1.15 | .20 |
| 15 | .5 | .4 | .43 | .45 | 40 | 1.0 | .4 | .95 | .44 |
| 16 | .5 | .6 | .57 | .57 | 41 | 1.0 | .6 | .95 | .52 |
| 17 | .5 | .8 | .53 | .63 | 42 | 1.0 | .8 | 1.02 | .86 |
| 18 | .5 | 1.0 | .48 | 1.20 | 43 | 1.0 | 1.0 | .84 | 1.04 |
| 19 | .5 | 1.2 | .48 | 1.40 | 44 | 1.0 | 1.2 | .55 | 1.74 |
| 20 | .5 | 1.4 | .30 | 2.28 | 45 | 1.0 | 1.4 | .71 | 1.78 |
| 21 | .5 | 1.6 | .40 | 1.95 | 46 | 1.0 | 1.6 | 1.07 | 1.49 |
| 22 | .5 | 1.8 | .61 | 1.80 | 47 | 1.0 | 1.8 | .84 | 1.89 |
| 23 | .5 | 2.0 | .52 | 2.01 | 48 | 1.0 | 2.0 | .76 | 1.97 |
| 24 | .5 | 2.2 | .60 | 2.10 | 49 | 1.0 | 2.2 | 1.35 | 1.86 |
| 25 | .5 | 2.4 | .00 | 14.78 | 50 | 1.0 | 2.4 | 1.17 | 1.86 |

TABLE 1--continued.

| Item number | a | b | $\hat{a}$ | $\hat{b}$ | Item number | a | b | $\hat{a}$ | $\hat{b}$ |
|---|---|---|---|---|---|---|---|---|---|
| 51 | 1.5 | -2.4 | 6.71 | -1.82 | 76 | 2.0 | -2.4 | * | -1.67 |
| 52 | 1.5 | -2.2 | 3.97 | -2.01 | 77 | 2.0 | -2.2 | * | -1.73 |
| 53 | 1.5 | -2.0 | 2.31 | -1.76 | 78 | 2.0 | -2.0 | * | -1.74 |
| 54 | 1.5 | -1.8 | 1.87 | -1.47 | 79 | 2.0 | -1.8 | 3.76 | -1.67 |
| 55 | 1.5 | -1.6 | 1.43 | -1.48 | 80 | 2.0 | -1.6 | 1.91 | -1.56 |
| 56 | 1.5 | -1.4 | 1.52 | -1.37 | 81 | 2.0 | -1.4 | 1.44 | -1.46 |
| 57 | 1.5 | -1.2 | 1.33 | -1.09 | 82 | 2.0 | -1.2 | 2.18 | -1.12 |
| 58 | 1.5 | -1.0 | 1.44 | -1.02 | 83 | 2.0 | -1.0 | 1.48 | - .99 |
| 59 | 1.5 | - .8 | 1.85 | - .73 | 84 | 2.0 | - .8 | 2.40 | - .63 |
| 60 | 1.5 | - .6 | 1.23 | - .61 | 85 | 2.0 | - .6 | 1.52 | - .61 |
| 61 | 1.5 | - .4 | 1.36 | - .48 | 86 | 2.0 | - .4 | 2.03 | - .27 |
| 62 | 1.5 | - .2 | 1.69 | - .13 | 87 | 2.0 | - .2 | 1.70 | - .18 |
| 63 | 1.5 | 0.0 | 1.19 | .08 | 88 | 2.0 | 0.0 | 2.13 | - .01 |
| 64 | 1.5 | .2 | 1.42 | .27 | 89 | 2.0 | .2 | 1.16 | .16 |
| 65 | 1.5 | .4 | 1.02 | .46 | 90 | 2.0 | .4 | 1.76 | .43 |
| 66 | 1.5 | .6 | 1.64 | .63 | 91 | 2.0 | .6 | 1.68 | .60 |
| 67 | 1.5 | .8 | 1.13 | .85 | 92 | 2.0 | .8 | 3.03 | .89 |
| 68 | 1.5 | 1.0 | .85 | 1.13 | 93 | 2.0 | 1.0 | 1.38 | 1.10 |
| 69 | 1.5 | 1.2 | * | 1.24 | 94 | 2.0 | 1.2 | 2.88 | 1.23 |
| 70 | 1.5 | 1.4 | .86 | 1.72 | 95 | 2.0 | 1.4 | * | 1.10 |
| 71 | 1.5 | 1.6 | .90 | 1.77 | 96 | 2.0 | 1.6 | * | 1.58 |
| 72 | 1.5 | 1.8 | .57 | 2.38 | 97 | 2.0 | 1.8 | * | .90 |
| 73 | 1.5 | 2.0 | * | 1.24 | 98 | 2.0 | 2.0 | .99 | 2.47 |
| 74 | 1.5 | 2.2 | .83 | 2.14 | 99 | 2.0 | 2.2 | * | .44 |
| 75 | 1.5 | 2.4 | 1.24 | 2.45 | 100 | 2.0 | 2.4 | 1.00 | 2.25 |

*Could not be estimated.

considered, the correlations between true and estimated parameter values were .85 and .99 for discrimination and difficulty, respectively.

The estimated Birn-3 parameter values for the 39 items in each of the six vocabulary tests are given in Table 2. In contrast to the Monte-Carlo items, parameter values were obtainable for all the vocabulary items. This appears to be due principally to the more restricted range of item difficulties in the vocabulary tests, although possibly the larger sample sizes also play a role.

Of the 234 items in Table 2, 175 have discriminatory parameter estimates of .8 or above and 139 of these are over 1.0. Urry (1970) has demonstrated that if the discriminatory power of the items is greater than .8, tailored testing is superior to conventional testing. According to the model, the discriminatory parameter values should be constant across populations. Since the Monte-Carlo data demonstrated that the estimation method is fairly accurate, especially in the -1.0 to +1.0 difficulty range, we have evidence that good tailored testing-item banks may be possible by combining conventional tests.

## 4. Discussion

One advantage in using Monte-Carlo data to study the estimation method presented in this paper was that guessing parameter values were known and did not have to be estimated with Lord's "eyeball" technique. With real data, the accuracy of item parameter estimates for both the maximum likelihood method and the method outlined here is severely restricted by the problems of estimating the guessing parameter values. The "eyeball" technique works satisfactorily only if the number of

TABLE 2

Birn-3 Parameter Estimates for Six Tests

| Item number | CTBS Level 2 | | | CTBS Level 3 | | | CTBS Level 4 | | | CAT Level 3 | | | CAT Level 4 | | | CAT Level 5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | a | b | c | a | b | c | a | b | c | a | b | c | a | b | c | a | b | c |
| 1 | 2.57 | -1.82 | .25 | .65 | -2.47 | .25 | .59 | -3.04 | .25 | .70 | -2.18 | .25 | 1.03 | -1.36 | .25 | .90 | -2.00 | .25 |
| 2 | 3.69 | -1.79 | .25 | 1.66 | -1.07 | .25 | .74 | -1.64 | .25 | 2.07 | -1.22 | .25 | 3.19 | - .80 | .25 | 1.10 | -1.58 | .25 |
| 3 | 1.63 | -1.47 | .25 | .74 | -1.64 | .25 | .69 | -1.62 | .25 | 1.02 | -1.14 | .25 | 1.46 | -1.74 | .25 | 1.24 | -1.50 | .25 |
| 4 | 1.11 | -1.17 | .25 | 1.27 | -1.32 | .25 | .94 | -1.06 | .25 | 2.90 | -1.09 | .12 | 1.21 | -1.19 | .25 | 1.08 | -1.33 | .25 |
| 5 | 1.58 | -1.30 | .25 | .76 | -1.20 | .25 | 1.26 | -1.04 | .25 | .60 | -1.31 | .25 | 1.40 | - .89 | .25 | 1.02 | -1.14 | .25 |
| 6 | .80 | -1.23 | .25 | 1.49 | - .72 | .25 | 1.31 | - .70 | .25 | 1.40 | - .53 | .20 | 2.21 | - .71 | .20 | 1.06 | - .93 | .25 |
| 7 | 2.31 | -1.46 | .25 | 2.17 | - .72 | .15 | 1.87 | - .92 | .25 | 1.25 | - .92 | .25 | 2.46 | - .70 | .20 | .67 | -1.37 | .25 |
| 8 | 1.54 | -1.16 | .25 | .99 | - .85 | .25 | 1.51 | - .76 | .25 | 1.66 | -1.14 | .25 | 2.70 | - .82 | .25 | .56 | -1.77 | .25 |
| 9 | .48 | -1.47 | .25 | 1.06 | - .66 | .25 | .30 | - .16 | .25 | 1.45 | - .93 | .25 | 1.54 | - .91 | .25 | .78 | -1.10 | .25 |
| 10 | .77 | -1.26 | .25 | .96 | - .34 | .25 | 1.06 | - .71 | .25 | 1.06 | - .66 | .15 | 1.45 | - .82 | .25 | .54 | -1.61 | .25 |
| 11 | 2.23 | -1.28 | .25 | 1.45 | - .93 | .25 | .54 | - .57 | .25 | 1.66 | -1.14 | .25 | 1.10 | - .46 | .25 | 1.17 | - .86 | .23 |
| 12 | .47 | - .88 | .25 | 1.72 | - .43 | .25 | .93 | - .76 | .25 | 1.05 | - .42 | .25 | 1.00 | -1.09 | .25 | .47 | .56 | .25 |
| 13 | 1.38 | - .85 | .20 | 1.54 | -1.16 | .25 | .80 | -1.02 | .25 | 1.23 | - .67 | .25 | 2.36 | - .74 | .25 | 1.29 | - .44 | .19 |
| 14 | 1.02 | - .65 | .23 | 1.11 | - .63 | .23 | .96 | - .69 | .25 | .83 | - .26 | .16 | 1.78 | - .27 | .18 | 1.51 | - .37 | .25 |
| 15 | .72 | -1.17 | .25 | 1.42 | - .37 | .25 | 1.04 | - .57 | .22 | 1.41 | - .51 | .18 | 1.29 | .97 | .25 | .71 | - .52 | .25 |
| 16 | .72 | - .70 | .25 | 1.33 | - .10 | .23 | 1.57 | .13 | .25 | .54 | - .57 | .25 | .98 | -1.03 | .25 | .96 | - .39 | .25 |
| 17 | 1.94 | -1.13 | .15 | 1.39 | - .32 | .18 | .47 | - .96 | .25 | 1.60 | - .28 | .15 | 1.22 | - .57 | .16 | 1.18 | - .73 | .25 |
| 18 | .29 | -2.47 | .25 | .79 | - .25 | .15 | 1.00 | - .43 | .25 | 3.42 | - .57 | .22 | 2.06 | - .53 | .25 | .81 | - .38 | .25 |
| 19 | .94 | -1.34 | .25 | .76 | .16 | .17 | .75 | - .29 | .25 | 1.39 | - .32 | .18 | 1.71 | - .56 | .18 | .97 | - .21 | .18 |
| 20 | .96 | - .51 | .18 | 1.86 | - .31 | .25 | 1.41 | - .10 | .25 | 1.59 | .02 | .15 | 1.15 | - .59 | .25 | .65 | .20 | .25 |

TABLE 2--continued.

| Item number | CTBS | | | | | | | | | CAT | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Level 2 | | | Level 3 | | | Level 4 | | | Level 3 | | | Level 4 | | | Level 5 | | |
| | a | b | c | a | b | c | a | b | c | a | b | c | a | b | c | a | b | c |
| 21 | .74 | - .51 | .25 | 1.12 | - .36 | .15 | 1.63 | - .24 | .25 | 1.22 | - .36 | .19 | 1.80 | - .48 | .18 | .93 | - .26 | .25 |
| 22 | .92 | - .23 | .17 | 1.29 | - .65 | .25 | 1.15 | - .59 | .25 | 1.49 | - .50 | .18 | 1.93 | - .23 | .20 | .89 | .21 | .25 |
| 23 | .92 | - .02 | .25 | .65 | .75 | .25 | 1.47 | - .27 | .16 | 1.52 | .18 | .20 | 1.72 | - .33 | .21 | 1.11 | - .12 | .20 |
| 24 | 1.13 | .12 | .22 | 1.44 | - .03 | .16 | 1.15 | - .66 | .17 | 1.48 | - .13 | .25 | 1.88 | - .27 | .25 | .89 | - .26 | .25 |
| 25 | 1.26 | - .10 | .25 | 1.11 | - .15 | .16 | .60 | - .05 | .16 | 1.16 | - .32 | .25 | 1.43 | - .29 | .25 | 1.11 | - .02 | .25 |
| 26 | .43 | .12 | .25 | .79 | .89 | .25 | .87 | - .02 | .13 | 1.05 | - .07 | .25 | 2.06 | - .35 | .16 | .54 | .10 | .25 |
| 27 | .92 | - .45 | .25 | 1.12 | .28 | .25 | .52 | .37 | .23 | .61 | .52 | .25 | .94 | .07 | .25 | 1.40 | - .05 | .15 |
| 28 | 1.47 | -1.05 | .25 | .70 | .93 | .20 | 1.01 | .29 | .25 | 1.55 | - .30 | .16 | .63 | .64 | .25 | 1.32 | .11 | .22 |
| 29 | .63 | - .20 | .12 | 1.07 | - .19 | .25 | .33 | 1.79 | .25 | .90 | .12 | .14 | 1.33 | .26 | .20 | .78 | .66 | .25 |
| 30 | .74 | .35 | .25 | 1.62 | - .06 | .25 | 1.13 | - .28 | .15 | 1.04 | .16 | .20 | .78 | .66 | .25 | 1.46 | .32 | .15 |
| 31 | .93 | .26 | .20 | 1.50 | .04 | .20 | 1.27 | - .52 | .25 | 1.27 | .27 | .22 | 1.25 | .54 | .18 | .88 | .37 | .22 |
| 32 | 1.28 | - .56 | .25 | .76 | 1.42 | .20 | .42 | 1.15 | .25 | 1.44 | .21 | .18 | .53 | 1.02 | .25 | 1.41 | .67 | .22 |
| 33 | 1.21 | .19 | .18 | .64 | .69 | .25 | 1.28 | .10 | .25 | .71 | .54 | .16 | .83 | 1.60 | .20 | 1.18 | .82 | .18 |
| 34 | 1.27 | - .81 | .25 | 1.52 | .49 | .25 | 1.34 | .27 | .12 | 1.66 | .10 | .20 | .79 | 1.06 | .19 | 1.60 | .87 | .14 |
| 35 | .57 | 1.04 | .25 | 1.21 | .22 | .16 | 1.31 | .97 | .25 | .85 | .40 | .18 | .84 | .96 | .19 | 1.24 | .79 | .23 |
| 36 | 1.04 | - .46 | .20 | .70 | .90 | .25 | .83 | .81 | .18 | 1.35 | .38 | .25 | .36 | 3.51 | .23 | 1.08 | .66 | .15 |
| 37 | .85 | 1.66 | .20 | .77 | .13 | .25 | 1.47 | .13 | .25 | 1.40 | .26 | .15 | .65 | 1.59 | .25 | .71 | 1.37 | .18 |
| 38 | .79 | .89 | .15 | .85 | .79 | .12 | 1.30 | .92 | .20 | .94 | .73 | .20 | .85 | 2.13 | .20 | 1.29 | .95 | .22 |
| 39 | .45 | .64 | .18 | 1.32 | .45 | .15 | .70 | 1.54 | .18 | .61 | 1.66 | .25 | .92 | .93 | .22 | 1.02 | 1.40 | .16 |

examinees is large. If the sample size is even moderate, there may simply not be enough low ability examinees to give an accurate estimate of the lower tail of the ICC. In this study, even with a sample size of 5,400, the estimation of $c_i$ values for the vocabulary tests involved considerable guesswork. Guessing parameter estimation must be improved before tailored testing techniques with multiple-choice items can become generally useful.

An examination of the equations used in the present parameter estimation method suggests two ways in which inaccurate $c_i$ estimates may cause problems. First, if the percentage of examinees passing an item is less than the estimate of $c_i$, (4) will yield a negative value. If this happens, the researcher's estimate of $c_i$ should obviously be reconsidered. The second problem is more complex. Under certain circumstances, the estimate of $\rho_{I\theta}$ in (15) will be greater than 1.0., e.g., when the $c_i$ parameter is overestimated. In (4) the overestimation of $c_i$ inflates the estimate of $P_i(\theta)$, which causes $\gamma_i$ in (7) and $\phi(\gamma_i)$ in (8) to be too small. The small value of $\phi(\gamma_i)$ and $(1 - c_i)$ in (15) may cause $\rho_{I\theta}$ to exceed 1.0. It should also be noted that, in general, overestimation of $c_i$ causes overestimation of both $a_i$ and $b_i$.

The Monte-Carlo data demonstrated that the estimation of item difficulty ($b_i$) by the method presented in this paper is accurate, but that estimation of item discrimination ($a_i$) can be relied on only for items with difficulties between -1.0 and +1.0. This places some limits on the usefulness of the method. The accuracy of the present method, compared with the maximum likelihood method, has not been completely investigated. The Urry graphic method, to which the present method is directly related,

has been shown (Jensema, 1972) capable of yielding estimates of $a_i$ and $b_i$ which correlate about .89 and .97 with maximum likelihood estimates. Until more extensive studies have been carried out, the method should be used with caution and must be viewed not as a replacement for maximum likelihood estimation, but as a convenient technique to economically prescreen items for further analysis.

The selection criteria used in prescreening the vocabulary items for a potential item bank were based on a number of factors. First, a good item bank contains items having a uniform distribution of difficulty and discrimination powers of at least .8. The six vocabulary tests covered a wide range of difficulty, and selecting roughly the same number of items from each test could be expected to yield a uniform distribution of difficulty over the range covered by the final item bank. Finally, it is known that the estimation method works best in the -1.0 to +1.0 range of difficulty and that the $c_i$ estimates may be somewhat inaccurate. After consideration of all these factors, a reasonable item selection criterion was to select all items in Table 2 that had difficulties between -1.2 and +1.2 and whose discrimination parameters were at least .8. There were 157 items which met these criteria.

The next step, which I intend to carry out and report in the future, is to administer the 157-item potential item bank to a large sample of examinees at the junior high school level. The responses of these examinees will then be reevaluated with maximum likelihood methods. Items that seem unsatisfactory will be discarded, and those remaining should form an excellent tailored testing-item bank.

Now, let us consider a few of the practical advantages of the estimation and prescreening methods outlined in this paper. First, remember that maximum likelihood methods are expensive. Each item that can be discarded without maximum likelihood parameter estimation saves several dollars worth of computer time, and the elimination of 82 items from the original 239 will result in quite a saving. Second, the elimination of 82 items also means a considerable reduction in testing time when the items are administered to a validation sample.

While it is true that some good items may have been discarded by the selection criterion used, the original pool of 239 items was large enough so that this was not important. Enough good items remain to create a good item bank. If saving every possible good item had been important, we could have modified the selection criteria and risked including a few bad items.

## REFERENCES


Birnbaum, A. Effective design and use of tests of mental ability for various decision making problems. Series Report No. 58-16. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, 1957. (a)

Birnbaum, A. Further considerations of efficiency in tests of mental ability. Technical Report No. 17, Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, 1957. (b)

Birnbaum, A. On the estimation of mental ability. Series Report No. 15. USAF School of Aviation Medicine, Randolph Air Force Base, Texas, 1957. (c)

Birnbaum A. Part 5: Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Jensema, C. J. An application of latent trait mental test theory to the Washington Pre-College Testing Battery. Research Bulletin. Seattle: University of Washington, Bureau of Testing, 1972.

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

Lord, F. M. An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. Psychometrika, 1953, 18, 57-76.

Lord, F. M. A note on the normal ogive or logistic curve in item analysis. Research Bulletin 64-56, Princeton: Educational Testing Service, 1964.

Lord, F. M. An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 1968, 28, 989-1020.

Lord, F. M. Item characteristic curves estimated without knowledge of their mathematical form--a confrontation of Birnbaum's logistic model. Psychometrika, 1970, 35, 43-50. (a)

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

Lord, F. M. Robbins-Monro procedures for tailored testing. Educational and Psychological Measurement, 1971, 31, 3-31.

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

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

Urry, V. W. A Monte-Carlo investigation of logistic mental test models. Unpublished doctoral dissertation, Purdue University, 1970.

Urry, V. W. Approximation methods for the item parameters of mental test models. Research Bulletin 0871-202. Seattle: University of Washington, Bureau of Testing, 1971. (a)

Urry, V. W. Individualized testing by Bayesian estimation. Research Bulletin 0171-177. Seattle: University of Washington, Bureau of Testing, 1971. (b)

## FOOTNOTES

[1]This research was partially supported by Office of Education Grant OEG-0-70-4797(607).

[2]The author wishes to express his appreciation to William M. Meredith for reading and commenting on an early draft of this paper.