program, including a tally of user errors by type.

In a sense, the evolution of these routines is coming full circle. Because of the pricing structure on TSS, which favors batch operations, and because of the large volume of conversational use during prime hours, several users have begun using the analysis of variance program in batch operations. The commands are arbitrary and meaningless in a batch mode without the attendant comments (2/3, 5/Y, etc.). Thus, a "final" version of the program, to be installed in late 1973, includes the option of a batch command set which is meaningful by itself. The input for the analysis shown in Fig. 1, for example, would be:

NDIM 2 LEVELS 3 5 DATA
[data for 15 cells as organized in "DUMMY" in Fig. 1.]
CONTRAST B −2 −1 0 1 2 STOP

Behavior Research Methods & Instrumentation
1974, Vol. 6, No. 2, 208-212

An interactive computer program for tailored testing based on the one-parameter logistic model

MARK D. RECKASE*
University of Missouri, Columbia, Missouri 65201

A program to implement tailored testing using the Rasch one-parameter logistic model is described and the problems encountered in its writing are discussed.

The concept of tailored testing (the administration to an individual of a specific set of items selected as appropriate) has been given considerable attention in the last 10 years in an attempt to overcome the problems inherent in the traditional objective testing situation. These problems concern time pressures on the individual, inefficient use of examinee time, inefficient use of item pools, lack of feedback, lack of objective success criteria, and many others. Reviews of the literature by Weiss and Betz (1973) and Cleary, Linn, and Rock (1968) discuss these problems in detail, so they will not be discussed here. The purpose of this paper is to present a computer program as one of the possible ways of solving these problems and to discuss the decisions that needed to be made in writing this program in the areas of (1) item selection, (2) scoring, (3) classification criteria, and (4) dimensionality of the item pool.

The theoretical base for the program is the one-parameter logistic model, commonly called the Rasch model (Rasch, 1960). This model relates the probability that individuals will answer item i correctly with a particular function of an individual's ability parameter and an item's easiness parameter. The relationship is given by the following formula:

$$P\{X_{si}\} = \frac{(A_s E_i)^{X_{si}}}{1 + A_s E_i}, \quad X_{si} = 0, 1$$

where $X_{si} = 0$ if the item was answered incorrectly and 1 if it was answered correctly, A_s is a parameter indicating person s's ability, and E_i is the easiness of item i. This model is a special case of the general three-parameter logistic model (Birnbaum, 1968):

$$P\{X_{si} = 1\} = c_i + \frac{(1 - c_i)e^{a_i(\theta_s - b_i)}}{1 + e^{a_i(\theta_s - b_i)}}$$

when $A_s = e^{\theta_s}$, $E_i = e^{-b_i}$, $a_i = 1$, and $c_i = 0$. In the three-parameter model, c_i is the guessing parameter for the item, a_i is the discrimination parameter, b_i is the difficulty parameter, and θ_s is the ability parameter for individual s.

The simpler Rasch model has been chosen for this program for several reasons. First, using the model, the ability parameters and item easiness parameters can be estimated independently (Rasch, 1960). This fact allows ability to be estimated on the same scale regardless of the set of items that is administered and allows items to be calibrated on groups at any ability level. The result of these estimation procedures is that item calibration can be performed conveniently on whatever groups happen to be available.

The second reason for using the one-parameter model is that efficient estimation procedures are available for estimating the parameters (Wright & Panchepakesan, 1969). The estimation procedures for the three-parameter model have so far been found to be much too lengthy for extensive use (Lord, 1968).

Finally, the one-parameter model is a special case of the exponential family of distributions which has the properties that the number of correct responses obtained

*Requests for reprints should be sent to Mark D. Reckase, 8 Hill Hall, University of Missouri, Columbia, Missouri 65201.

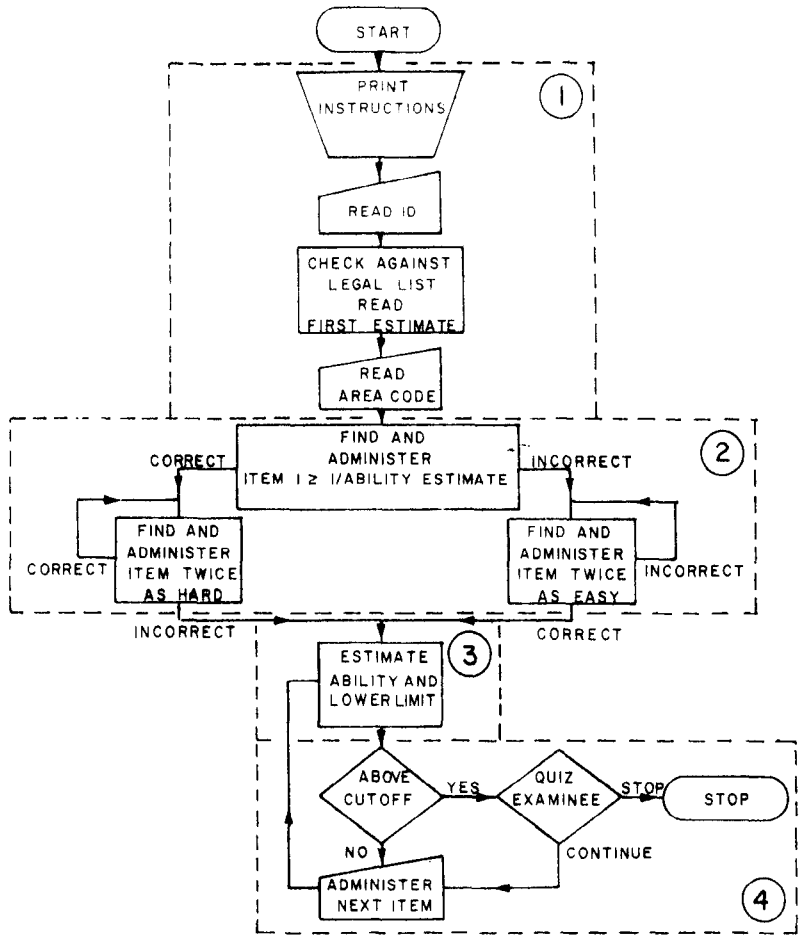

Fig. 1. Program flow chart.

by a person is a sufficient statistic for the ability parameter and the number of correct responses by a number of individuals to an item is a sufficient statistic for the easiness parameter. These conditions make the Rasch model particularly convenient for tailored testing.

However, in exchange for the conveniences, three somewhat restrictive assumptions have to be made. First, the model assumes that the probability of obtaining a correct response by guessing is insignificant. Since multiple-choice items are used, this assumption is obviously not met, especially for Ss of low ability. However, Ross (1966) found that guessing had little effect on the fit of the Rasch model. Second, the model assumes that all items are of equal discrimination. This assumption can easily be met if items are chosen carefully, but it does restrict the size of the item pool and, in practice, items of varying discrimination are used. The third assumption is that the trait being measured is unidimensional. Thus, the tailored tests must be restricted to homogeneous subject areas.

Since the restrictions placed by the assumptions do not seem to outweigh the conveniences of the method, the one-parameter model has been chosen. However, robustness studies are currently being undertaken to assess the consequences of violating the assumptions.

DESCRIPTION OF THE PROGRAM

The program for implementing tailored testing using the one-parameter logistic model is written in FORTRAN IV for use on an IBM System 370/165 computer with time-sharing option (IBM, 1972). The program also assumes that an interactive terminal is available for presentation of instructions and test items and for input of response. An IBM 2741 typewriter

```
TERMINAL TESTING PROCEDURE
YOU WILL BE PRESENTED WITH A SERIES OF TEST ITEMS.  RESPOND TO
EACH ITEM BY TYPING THE APPROPRIATE LETTER AND PRESSING THE
RETURN KEY.  ITEMS WILL BE PRESENTED UNTIL A CLEAR DECISION
IS REACHED CONCERNING WHETHER YOU ARE ABOVE OR BELOW A C GRADE.
IF YOU WISH TO CONTINUE ON FOR A HIGHER GRADE, INSTRUCTIONS
WILL BE GIVEN AT THAT POINT.  IF AT ANY TIME YOU WISH TO STOP
BEFORE A DECISION HAS BEEN MADE, TYPE THE WORD  STOP  AFTER
YOUR LETTER RESPONSE AND PRESS THE RETURN KEY.

PLEASE TYPE YOUR STUDENT NUMBER AND PRESS THE RETURN KEY
IF YOUR STUDENT NUMBER CONTAINS ONLY 5 DIGITS, START IT
WITH A LEADING ZERO TO MAKE 6 DIGITS.


100000
INPUT:    ID = 100000
TYPE THE CODE CORRESPONDING TO THE AREA YOU ARE TO BE TESTED ON
          SM FOR STATISTICS AND MEASUREMENT
          ET FOR CLASSROOM EVALUATION TECHNIQUES
          ST FOR STANDARDIZED TESTS
AFTER TYPING THE PROPER CODE, PRESS THE RETURN KEY
sm
INPUT:    TEST CODE = SM
```

Fig. 2. Instructions to examinee.

```
        WHICH OF THE FOLLOWING SETS OF STATISTICAL DATA HAS MOST LIKELY
        BEEN MISCALCULATED?
             (A)  RANGE = 1-50;  MEAN = 25;  S.D. = 12
             (B)  RANGE = 10-100;  MEAN = 75;  S.D. = 15
             (C)  RANGE = 15-20;  MEAN = 18;  S.D. = 7
TYPE RESPONSE LETTER AND PRESS RETURN


c
CORRECT

        A HIGH CORRELATION MAY EXIST BETWEEN TWO VARIABLES
             (A)  ONLY WHEN THE VARIANCES OF THE TWO VARIABLES ARE
                  FAIRLY CLOSE TO BEING EQUAL.
             (B)  ONLY WHEN THE MEANS OF THE TWO VARIABLES ARE FAIRLY
                  CLOSE TO BEING EQUAL.
             (C)  EVEN THOUGH THE MEANS AND VARIANCES OF THE TWO
                  VARIABLES ARE EXTREMELY DIFFERENT.
TYPE RESPONSE LETTER AND PRESS RETURN


b
INCORRECT

ABILITY ESTIMATE =    1.198
LOWER LIMIT OF ABILITY ESTIMATE =    0.57
```

Fig. 3. Sample item administration.

terminal is currently being used with this program, but cathode ray terminals are preferable since they do not give a printed copy of the test items.

The program is composed of a main program and one subroutine which performs the item search. Three data sets are also required for the operation of the program; one contains the calibrated test items, the second contains record data on the Ss tested, and the third, which is optional, receives information on the operation of the program. Logical operation of the program is divided into four major segments. These segments are enclosed within the broken lines on the flow chart shown in Fig. 1.

The first program segment performs the setup operations for the program (see Fig. 2). These operations include printing instructions, reading an examinee ID and checking it against a list of legal examinees, reading initial ability estimates from the examinee record data set, and reading a code that tells which of the possible sets of test items and subareas within a set of test items should be administered. This last feature will be discussed further in a later section of the paper. It has been included because of the unidimensional assumption of the Rasch model.

The second program segment administers items to the examinee until his ability can be estimated (see Fig. 3). An estimate cannot be obtained until both a correct and an incorrect response have been recorded. To efficiently arrive at a point where the examinee's ability can be estimated, an item with easiness equal to the reciprocal of the original estimate of the examinee's ability is searched for and the item with easiness equal to or greater than that desired value is administered. An item with easiness equal to the reciprocal of ability has a traditional difficulty value of 50% for the examinee. If no original estimate of ability is available, the first item administered will have easiness equal to 1.0.

If the examinee responds correctly to the first item, an item with one-half the easiness is administered and the halving procedure continues until an incorrect response is obtained. If the first item is answered incorrectly, an item with twice the easiness is administered and the doubling procedure is continued until a correct response is obtained. Once both correct and incorrect responses are present in the response vector, the S's ability can be estimated and the third segment of the program is entered.

The third segment of the program estimates the ability parameter of the S using an iterative maximum likelihood algorithm and then computes a lower limit on the ability estimate from the normalized likelihood distribution. After the lower limit is determined, the program continues on to the fourth and final segment.

The last program segment compares the lower limit on the ability estimate to predetermined cutoff values labeled C, B, and A. If the lower limit exceeds a cutoff value, the examinee is quizzed by the terminal to determine if he wishes to go on to a higher level. If not, the session is terminated. The session is also terminated if the A-level cutoff is exceeded by the lower bound. If the session is not terminated, an item with easiness equal to the reciprocal of the estimated ability is searched for and the nearest item equal to or easier than that desired is administered. After the item is responded to, Segment 3 of the program is reentered and the procedure continues as described above.

Once the program begins the cycle of estimating ability, searching for an appropriate item, administering the item, estimating ability, etc., the program may be terminated in four ways. First, at any time the S may type "STOP" after his response and the program will terminate. Second, after exceeding a cutoff, a termination opportunity is presented and the S may discontinue processing. Third, if the A cutoff is exceeded, the program terminates automatically. And finally, if the program runs out of items, the program terminates.

INPUT DATA

In order for the program to operate properly, three input sources are required. The first is the terminal itself,

```
07780  25320CSMCT00000 1916 5 373
07800  25320THE MEAN, MEDIAN, AND MODE ARE ALL
07820  25320      (A) INFERENTIAL STATISTICS.
07840  25320      (B) MEASURES OF VARIABILITY OF SCORES.
07860  25320      (C) MEASURES OF CENTRAL TENDENCY.
07880  25320      (D) TYPES OF PERCENTILLS.
07900  32490DSMCT00000 2344 5 373
07920  32490WHAT IS THE MEAN OF THE FOLLOWING SCORES: 2, 5, 3, 9, 8, 7?
07940  32490      (A) 4
07960  32490      (B) 6                          X̄= (SUM X)/N
07980  32490      (C) 3
08000  32490      (D) 5
```

Fig. 4. Representation of items in item data set.

which asks the S for his ID code, subject area to be tested, stop codes, and responses to the test items. Input from the terminal is read by the program in alphanumeric form to allow checking for correctness.

The second input source is an examinee record file which contains legal ID codes, ability estimates for each of the possible test areas, and counts of the number of test sessions in each area. Special codes are also included to control the number of test sessions allowed during certain time periods. The ability estimates are used as a starting point for determining the items to administer.

The third input data set contains the item pool along with calibration data. All of the items included in this data had first been calibrated using a modification of a program described by Wright and Panchepakesan (1969) using groups varying from 170 to 960 Ss. All items were placed on the same easiness scale by placing reference items in each test and determining a multiplicative scaling constant on the basis of these items. The items currently calibrated are from an item pool used with an introductory measurement course. Approximately 150 items in several different areas are available.

Each test item in the data set is stored with the following information: (1) easiness parameter based on total area test (e.g., statistics); (2) correct response; (3) code giving general area (e.g., SM for statistics and measurement); (4) a code for the specific area within the general area (e.g., CT for central tendency); (5) easiness parameter for the specific area; (6) number of lines required for the items; and (7) data of latest calibration. An example of an item in the data set is shown in Fig. 4. The numbers to the far left are line numbers used for updating purposes.

DISCUSSION

As mentioned earlier, decisions had to be made when this program was written concerning how items would be selected, how the procedure would be scored, how classification decisions could be made, and what should be done about the multidimensional nature of the item pool.

In this program, the item selection procedure and the scoring procedure are inseparable, so they will be discussed together. The scoring procedure consists of the maximum likelihood estimate of the ability parameter after each item is administered. This technique is quite different from the total correct, reciprocal of average easiness, or reciprocal of final easiness discussed by Lord (1968), in that the total response pattern is taken into account rather than just a summary statistic. The maximum likelihood estimate converges rather quickly to a stable value, as is shown in Fig. 5. Limited empirical testing has shown that adequate convergence seems to occur after about 10 items.

After each ability estimate, an item is selected that has easiness equal to the reciprocal of the ability. As the ability estimate converges, the easiness value also converges on an easiness that will give traditional difficulty of 50%. Thus, the difference between the easiness of the items administered decreases as the procedure continues. This yields a set of items similar to that determined by the Robbins-Monro process, which Lord (1970) has suggested might be a preferred method of selection.

The above discussion concerns the estimation of the ability parameter, but, as Green (1970) has suggested in his comment of Lord's (1970) article, the purpose of a test situation may not be strictly to estimate ability but may also involve the classification of S's into categories. In many cases, the classification may be performed with many fewer items than are required to obtain a good estimate of the ability parameter. Part of the function of this program is to classify Ss into categories above the C, B, or A cutoffs. Therefore, a technique had to be chosen to perform this classification. In this case, the likelihood distribution approximated for the purpose of estimating the ability parameter was conveniently available for use in setting limits on the ability estimate. The logic behind the procedure is that, if the response pattern obtained has a low probability for a given ability estimate, the estimate is most likely incorrect. In this

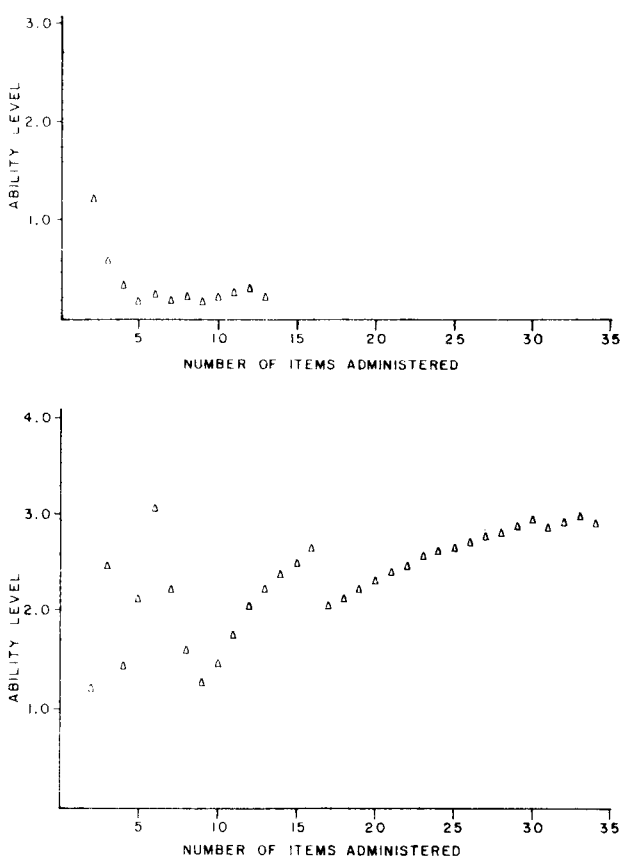

Fig. 5. Two examples of ability parameter convergence.

case, after a number of items have been administered, the response pattern has a low probability for any response estimate, so the important consideration is the probability relative to the maximum. The procedure eliminates those ability estimates from contention that are below the maximum likelihood estimate and yield the lowest probability for the obtained response pattern. The actual steps involve approximating the area beneath the likelihood distribution and then setting the lower limit of the ability parameter at the point that cuts off the lower 5% of the distribution. If this lower bound is above a cutoff value, the subject is classified. Classification above the C level has been performed with as few as four items.

The final consideration included in the program is that of the dimensionality of the item pool. The one-parameter logistic model assumes that the trait being measured is unidimensional, a situation that is obviously not the case when the item pool is made up of items dealing with standardized tests, statistics, and classroom tests. To overcome this problem, the items were divided into three general areas and each area was calibrated separately. The areas were then administered separately by specifying the proper code at the start of the test situation.

To further comply with the assumption, a second calibration was carried out with a greater restriction on the test area measured. For instance, all items on central tendency were grouped together and calibrated. The program allows the option of choosing to administer just items from these more restricted areas.

SUMMARY

A program to implement tailored testing using the Rasch one-parameter logistic model has been described and the problems encountered in its writing were discussed. The problems included determining the method of scoring, the item selection process, the classification criteria, and the multidimensional nature of the item pool. The solutions used in writing the program were arriving at an ability estimate using maximum likelihood as the scoring procedure, choosing the item with easiness greater than or equal to the reciprocal of the ability estimate as the item selection technique, setting the lower limit of ability for selection purposes at the lower 5% point of the likelihood distribution, and performing separate item calibrations on each content area to satisfy the dimensionality assumption.

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, Mass: Addison-Wesley, 1968. Chapters 17-20.

Cleary, T. A., Linn, R. L., & Rock, D. A. An exploratory study of programmed tests. Educational & Psychological Measurement, 1968, 28, 345-360.

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

International Business Machines. *IBM System 360 operating system time sharing option terminal user's guide.* Poughkeepsie, N.Y: IBM, 1972.

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

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.

Rasch, G. *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Denmarks Paedagogiske Institut, 1960.

Ross, J. An empirical study of a logistic mental test model. Psychometrika, 1966, 31, 325-340.

Weiss, D. J., & Betz, N. E. Ability measurement: Conventional or adaptive? Research Report 73-1, Psychometric Methods Program, Department of Psychology, University of Minnesota, February 1973.

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