DOCUMENT RESUME

ED 091 419 TM 003 633

AUTHOR	Kalisch, Stanley J.
TITLE	A Tailored Testing Model Employing the Beta Distribution and Conditional Difficulties.
PUB DATE	[74]
NOTE	20p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, Illinois, April, 1974)
EDRS PRICE	MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS	*Branching; Complexity Level; Computer Oriented Programs; Decision Making; Guessing (Tests); Item Sampling; *Models; Prediction; Probability; *Programed Materials; *Response Style (Tests); *Testing; Test Validity
IDENTIFIERS	Loss Function; *Tailored Testing; Variance (Statistical)

ABSTRACT

A tailored testing model employing the beta distribution, whose mean equals the difficulty of an item and whose variance is approximately equal to the sampling variance of the item difficulty, and employing conditional item difficulties, is proposed. The model provides a procedure by which a minimum number of items of a test, consisting of a set of pre-specified items, is presented to an individual, and the correctness of the individual's responses to the remaining items is predicted. A validation study of the procedure indicates that the model is feasible. (Author)

Session # 27.18

A TAILORED TESTING MODEL EMPLOYING THE BETA DISTRIBUTION AND CONDITIONAL DIFFICULTIES

Stanley J. Kalisch
Center for Educational Design
Room 8 Kellum Hall
Florida State University
Tallahassee, Fla. 32306

A paper submitted at the annual meeting of
The American Educational Research Association
Chicago, Illinois
April, 1974

ABSTRACT

A tailored testing model employing the beta distribution, whose mean equals the difficulty of an item and whose variance is approximately equal to the sampling variance of the item difficulty, and employing conditional item difficulties, is proposed. The model provides a procedure by which a minimum number of items of a test, consisting of a set of pre-specified items, is presented to an individual, and the correctness of the individual's responses to the remaining items is predicted. A validation study of the procedure indicates that the model is feasible.

INTRODUCTION

With the increased stress on individualized instruction, in which frequent measurement of student performance is necessary, a reduction in the amount of testing time with minimal loss of information is desirable. A reduction in testing time is also desirable in computer-based instruction as a means of diminishing on-line costs.

The purpose of this paper is to present a tailored testing model which may be used in an instructional setting for proper placement of individuals in an instructional sequence, and for evaluation of an individual's performance after instruction. The proposed testing model provides a means of reducing testing time by providing a procedure of presenting to an examinee a minimum number of items of a test, consisting of predetermined items, and predicting the correctness of the responses the examinee would have made if he had been presented with the remaining items. The model employs the beta distribution, item conditional difficulties, and an expected loss function. A validation study of the model was conducted by performing a computer simulation on existing data.

Research regarding tailored testing has primarily been conducted in the area of aptitude and achievement tests, with regard to the measurement of underlying traits and not with regard to instructional testing (Cleary, Linn, & Rock, 1968, 1968(a); Hubbard, 1966; Linn, Rock, & Cleary, 1969; Lord, 1971; Waters & Bayroff, 1971). Ferguson (1970) investigated an instructional tailored testing procedure involving a hierarchical arrangement of objectives. Lord (1970) defines a tailored test as one in which the presentation of items to an individual is determined by his responses to the previous items, as a means of attaining optimal measurement of the individual's ability. Cleary, Linn, and Rock (1968) define tailored tests as tests which contain a sequential branching system that presents to the individual items which are appropriate to his level of performance. Although Lord (1970) did not consider a tailored test theory for instructional tests, he differentiated between the purposes of testing in instructional and measurement situations. Lord contends that a measuring instrument should not alter the underlying trait being measured, whereas instructional tests are to measure an underlying trait which is to be, or which has been, altered by instruction. Green (1970) has indicated that a complete discussion of tailored testing should include consideration of the interplay between instruction and testing in computer-based situations.

The tailored testing model proposed in this paper was designed in terms of Cleary, Linn, and Rock's (1968) definition in which a branching system is used to present test items that are appropriate to the individual's level of performance. The model is applicable to instructional situations in which objectives and corresponding test items are employed. The purpose of the model is to reduce the number of test items an individual receives and obtain the same information concerning the correctness of each item response as would be determined if the total test had been administered to the individual. The model uses a branching technique based upon the examinee's responses to all previously presented items. The model is not dependent upon an assumed or validated hierarchy of skills.

THEORETICAL FRAMEWORK

Conditional Difficulty of an Item

The proposed model uses a branching strategy and a decision making procedure that require determination of the probability that an examinee will answer a specified item correctly. The item difficulty provides such a probability value, but fails to make use of the information concerning the correctness of the examinee's responses to the previously presented items. Use of such information, although impractical, if not impossible, in paper and pencil testing situations, is possible in computer-based situations. Hence, the model employs the use of item conditional difficulties.

The conditional difficulty of an item for an individual is the probability of the individual's answering the k-th item correctly, given information concerning the correctness of the individual's responses to the preceding k-1 items and given prior subject—item response data. Hence, a conditional difficulty is a conditional probability. Let C_i represent the event that item i was answered correctly; W_i represent the event that item i was answered incorrectly; and P(B|A) represent the probability of event B given event A. Figure 1 contains a subject by item matrix on ones and zeros representing correct and incorrect responses, respectively. If an individual has answered the first item correctly, the conditional difficulty of the remaining two items is determined as follows:

$$P(C_2|C_1) = \frac{P(C_2 \cap C_1)}{P(C_1)} = \frac{5}{8} \qquad [1]$$

$$P(C_3|C_1) = \frac{P(C_3 \cap C_1)}{P(C_1)} = \frac{4}{8} . \qquad [2]$$

SUBJECTS	ITEMS		
	1	2	3
1	1	1	0
2	1	1	1
3	1	1	0
4	1	1	1
5	1	1	0
6	1	0	1
7	0	1	0
8	1	0	1
9	0	0	0
10	1	0	0

Fig. 1. Subject by item response matrix for ten subjects and three items.

If the individual correctly answered the first item but incorrectly answered the second item, the conditional difficulty of the third item is

$$P(C_3|C_1,W_2) = \frac{P(C_3 \cap C_1 \cap W_2)}{P(C_1 \cap W_2)} = \frac{2}{3}. \qquad [3]$$

Item Difficulty and the Beta Distribution

The tailored testing model being proposed uses population item difficulties. Since the values obtained for the difficulty of an item may vary according to the samples of examinees selected, the model considers the sampling variance of the difficulty of an item. For any item with a sample difficulty value between zero and one, there exists a beta distribution which has the following properties: (1) the expected value of the distribution equals the obtained item difficulty; (2) its variance is nearly equal to the sampling variance of the difficulty of the item; and (3) its domain is the closed interval from zero to one. It has therefore been assumed that the beta distribution approximates the distribution of item difficulty values obtained from infinitely many samples of examinees.

Consider a dichotomously scored item answered by N individuals, r of whom answered the item correctly. An unbiased estimate of the population parameter of item difficulty is p = r/N. For an item (Hays & Winkler, 1970), the sample variance s^2 is given by the equation

$$s^2 = \frac{Nr-r^2}{N^2} = p(1-p) \; ; \qquad [4]$$

an unbiased estimate of the population variance $\hat{\sigma}^2$ is given by the equation

$$\hat{\sigma}^2 = \frac{N}{N-1}s^2 = \frac{Nr-r^2}{N(N-1)} = \frac{Np(1-p)}{N-1} \; ; \qquad [5]$$

and the variance of the sampling distribution of item difficulties σ_M^2 is given by the equation

$$\sigma_M^2 = \frac{\hat{\sigma}^2}{N} = \frac{r(N-r)}{N^2(N-1)} = \frac{p(1-p)}{N-1}. \qquad [6]$$

The previous discussion concerning item difficulty also applies to conditional item difficulty.

The beta distribution (Hays & Winkler, 1970), specified by the equation

$$f(p) = \begin{cases} \frac{(N-1)!}{(r-1)!(N-r-1)!} p^{r-1}(1-p)^{N-r-1} & \text{if } 0 \le p \le 1 \\ 0 & \text{if } p < 0 \text{ or } p > 1, \end{cases} \qquad [7]$$

defines the random variable p in terms of the probability of a "success" on any single Bernoulli trial. The mean of the distribution is r/N, and the variance of the beta distribution σ_B^2 is given by the equation

$$\sigma_B^2 = \frac{r(N-r)}{N^2(N+1)} = \frac{p(1-p)}{N+1} , \qquad [8]$$

The difference of σ_M^2 and σ_B^2 is calculated as follows:

$$\sigma_M^2 - \sigma_B^2 = \frac{p(1-p)}{N-1} - \frac{p(1-p)}{N+1} = \frac{2p(1-p)}{N^2-1}. \qquad [9]$$

For fixed values of N, the difference is greatest when $p = .5$; and for fixed values of p, not equal to zero or one, the difference diminishes as N increases. From equation 7, since $r \geq 1$ and $N-r-1 \geq 0$, $N \geq 2$. The largest value of expression 6 is approximately 0.167 (under the conditions that $N = 2$ and $p = .5$). If it is desired that the beta distribution more nearly approximate the assumed distribution of sample difficulties of an item, a requirement that $\sigma_M^2 - \sigma_B^2 \leq \epsilon$, for $0 < \epsilon < .167$, may be imposed. Such a restriction requires a minimum value of N, which is derived from expression 9 as follows:

$$\sigma_M^2 - \sigma_B^2 = \frac{2p(1-p)}{N^2-1} \leq \epsilon$$

$$\frac{2}{\epsilon}p(1-p) \leq N^2-1$$

$$N^2 \geq 1 + \frac{2p(1-p)}{\epsilon}$$

$$N \geq \sqrt{1 + \frac{2p(1-p)}{\epsilon}} . \qquad [10]$$

The shape of the beta distribution is dependent on r and N (Hays & Winkler, 1970). Figure 2 shows three distributions in which $r = N/2$. The distributions are symmetric with respect to the line with the equation $p = .5$. If $N = 2$ and $r = 1$, the distribution is rectangular.

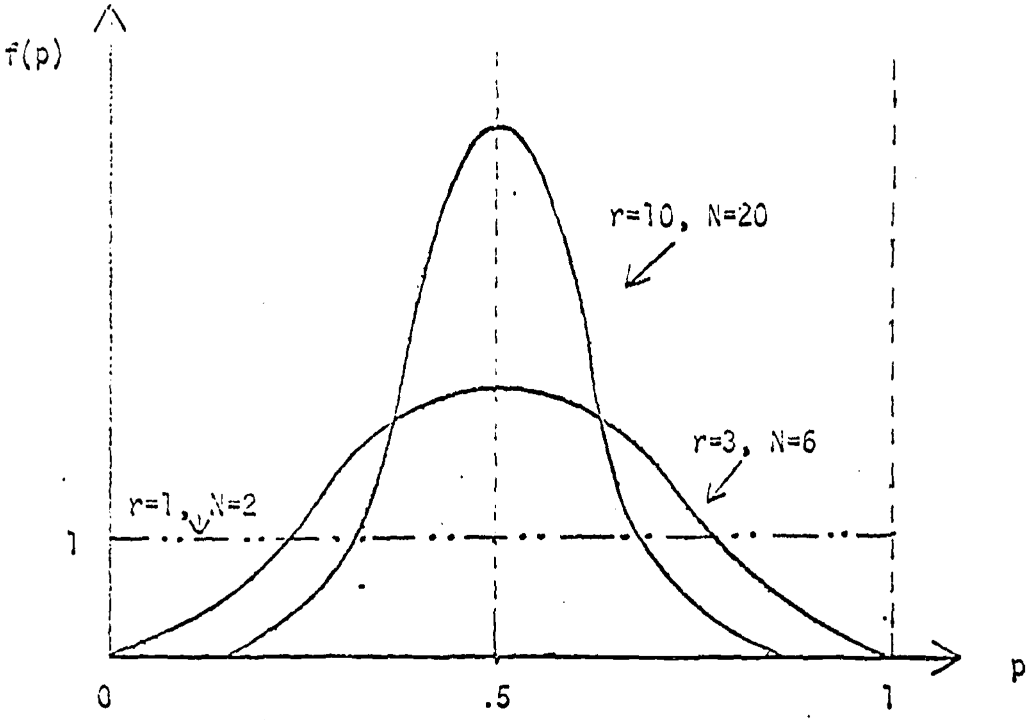

Fig. 2. Three symmetric beta distributions.

If r < N/2, the distribution is positively skewed; and if r > N/2, the distribution is negatively skewed. If r > 1 and N > 2, the distribution is unimodal with the mode equal to (r-1)/(N-2); and if N ≥ 2 and r = 1 or r = N-1, the distribution is unimodal with the mode at zero or one, respectively. Figure 3 shows three skewed, unimodal beta distributions.

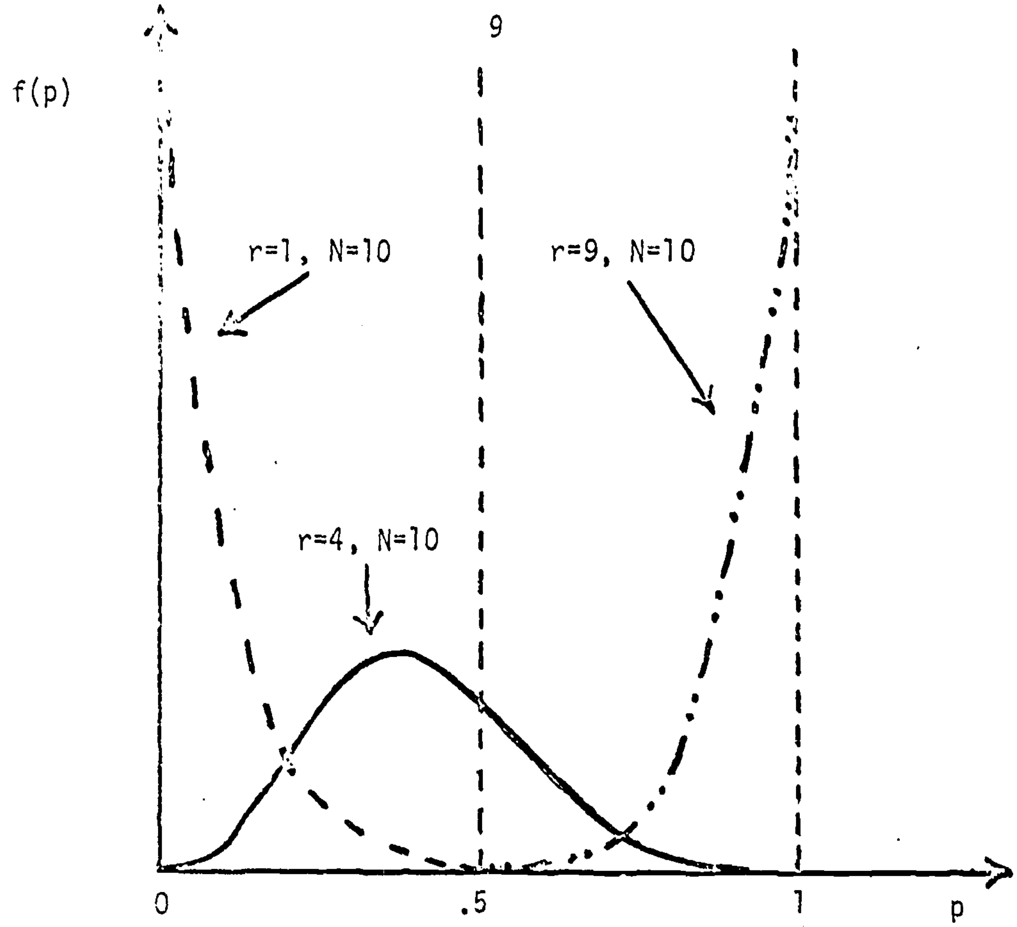

Fig. 3. Three skewed beta distributions.

The model being proposed uses a loss function weighted with the probabilities of the population difficulties of items being within specified intervals. Such probabilities may be determined by using the beta distribution. Since the area between the horizontal axis and the beta distribution curve is one unit, the probability that p is in the interval [a,b] with respect to the beta distribution is calculated from the formula

$$P(a \le p \le b) = \int_a^b f(p)dp$$

$$= \int_a^b \frac{(N-1)!}{(r-1)!(N-r-1)!} p^{r-1}(1-p)^{N-r-1} dp. \qquad [11]$$

Loss Functions

The tailored testing procedure being suggested requires the prediction of thc correctness of an examinee's responses to items, based upon his responses to previously presented items. If the population conditional difficulty of an item for an individual were high, such as greater than 0.9, the decision maker might be willing to assume that the examinee would answer the item correctly if it were presented to nim. A low conditional difficulty, such as less than 0.1, might result in a decision to assume the examinee's response would be incorrect. If the population difficulty were between 0.1 and 0.9, the decision maker might wish not to predict the correctness or incorrectness of the examinee's response. A condition of uncertainty exists since the actual value of the population conditional item difficulty is not known.

In situations in which an individual must make a decision under a condition of uncertainty, the consequences of a decision may be expressed in terms of a loss to the individual. The loss (Hays & Winkler, 1970) is the result of a combination of (1) the decision maker's action; and (2) the actual state of the world, or information categories (Cronbach & Gleser, 1965). Losses may be expressed in monetary units; however, certain factors are not directly or solely related to monetary units. The theory of utility provides a means of measuring the relative value of losses in a decision problem.

Figure 4 depicts a loss function applicable to the tailored testing model being proposed. The dimension, information categories, specifies three possible cases involving the population conditional difficulty ρ of an item: $\rho < .10$; $.10 \leq \rho \leq .90$; and $\rho > .90$. The second dimension, the action taken, is separated into three levels: (1) assume S's response would be correct; (2) make no assumption; and (3) assume the response would be incorrect.

ACTION	INFORMATION CATEGORIES		
	$\rho < .10$	$.10 \leq \rho \leq .90$	$\rho > .90$
Assume response would be correct	1000	100	1
Make no assumption	10	1	10
Assume response would be incorrect	1	100	500

Fig. 4. Loss function.

The nine hypothetical values in the matrix are the expected relative losses for a specified action given a particular information category. The values of the loss function may incorporate such factors as the following: (1) the additional cost of instruction; (2) the additional instructional time needed; and (3) the effect on the student's morale due to the improper placement of the student in an instructional sequence, on the basis of the predicted responses. In a computer-based testing situation, a greater loss may be attributed to the increased time and cost needed in presenting an examinee an item and processing his response than in predicting the correctness of his response. These relative losses may also be reflected in a loss function. The values of a loss function may be determined by consensus of the decision makers or from the results of previously obtained data. According to the loss function in Figure 4, if the population conditional difficulty ρ is less than 0.10, the expected loss in assuming an individual's response to be correct is 1000 times the expected loss if the decision is to assume the response to be incorrect.

Decision theory provides a means of weighting the levels within the state of the world. In terms of conditional probabilities, a point estimate $(\hat{\rho} < .10)$ of the population conditional difficulty ρ does not imply that the probability of $\rho < .10$ is one. If the probabilities for the levels for the information categories were .5, .4, and .1, the expected loss, denoted by EL, for each action would be computed as follows:

EL(Assume response would be correct)

$$= (.5)(1000)+(.4)(100)+(.1)(1) = 540.1$$

EL(Make no assumption)

$$= (.5)(10)+(.4)(1)+(.1)(10) = 6.4$$

EL(Assume response would be incorrect)

$$= (.5)(1)+(.4)(100)+(.1)(500) = 90.5.$$

According to decision theory, the minimum expected loss of 6.4 dictates that the best action is to make no assumption. The probabilities for the levels within the information categories, specified in Figure 4, may be computed by using formula 11.

THE MODEL

The tailored testing model proposed is specified by the following procedure:

1. A data base consisting of a subject by item matrix of dichotomously scored items, as illustrated in Figure 1, is obtained.
2. Critical difficulty values V_1 and V_2 are selected, such that if for any S the population conditional difficulty ρ is less than V_1, it is assumed that S would answer the item incorrectly; and, if ρ is greater than V_2, it is assumed that S would answer the item correctly.
3. A loss function, as exhibited in Figure 4, is specified.
4. A minimum number of items to be presented to each S, prior to any item response predictions for S, is specified.

5. The maximum difference permitted between σ_M^2 and σ_B^2 is specified. The minimum number of observations necessary for prediction is determined by formula 7, provided the value specified is in the open interval (0,.167). If no minimum difference is specified, then $N \geq 2$ is required for use with the beta distribution.
6. Responses from S to the minimum number of items indicated in Step 4 are obtained. The branching strategy used in the model for presentation of items requires the difficulty values of the first items presented and the conditional difficulties of the subsequently presented items be closest to 0.5. Other branching strategies might be employed, but the strategy of successive classification of S as a member of either one of two halves of a population is intuitively expedient if not also optimal. Each decision in the branching strategy is dependent on all the previous item responses given by the individual, not merely the immediately preceding response, as is done in Markov chain theory.
7. For each unpresented item for which no response prediction has been made, probabilities are computed for the three information categories: $\rho < V_1$; $V_1 \leq \rho \leq V_2$; and $\rho > V_2$. If the conditional difficulty of the item does not equal one or zero, the probabilities are calculated from formula 11. If the conditional difficulty equals one or zero, the sampling variance of the conditional difficulty is zero. In these two cases, a probability of one is assigned to the corresponding level of the information categories and zero to the other levels. Using the obtained probabilities and the specified loss function, the expected loss for each action is computed. If the minimum of the three values indicates prediction to be the best

decision, then the specific prediction is noted and the item is removed from further consideration.

8. For each unpresented item for which no prediction has been made, the conditional difficulty of the item is determined. The branching strategy is then employed. The item whose conditional difficulty is closest to 0.5 is then presented. After obtaining a response to the item, the procedure beginning with Step 7 is repeated.
9. As additional items are presented to S, the number of observations in the data base upon which conditional probabilities are computed is reduced. If the minimum number of observations is not available, the remaining items for which responses have neither been obtained nor predicted, are presented to S for his responses.

A VALIDATION OF THE MODEL

Method

A data base containing the item scores for each of 20 items and for each of 62 Ss was obtained. The test was administered as a pretest for prospective elementary school teachers enrolled in two sections of a mathematics course at Florida State University during the fall and winter quarters 1972-73.[1] The stem of each item was presented without the alternatives to each S. Upon completing his constructed responses to the items, the S was presented with the item stems and six alternatives for each stem. If S had failed to supply a constructed response for an item, he was to choose the alternative, "I did not answer the question." The letter of each

[1] Design and administration of the test were performed by Drs. Lee Armstrong and Ann Joyner under the direction of Dr. Robert Kalin of the Department of Mathematics Education at Florida State University.

alternative was entered by S at a computer terminal connected to the Control Data Corporation (CDC) 6500 computer at Florida State University. After entering the responses to all items, S was informed how many items he had answered correctly, and assignments corresponding to the items incorrectly answered were prescribed.

Using program TESTAT (Veldman, 1967), it was determined that the mean number of correct responses was 8.5 per S; the standard deviation was 5.7; the reliability of the test, using KR-20, was .92; and the difficulties of the items ranged from .21 to .73.

A validation group consisting of 38 students in the same mathematics course during the spring quarter was used. Ss in the validation group received the same test items, presented in the same order and with the same directions as the previous students. The critical difficulties selected were $V_1 = 0.10$ and $V_2 = 0.90$; the loss function in Figure 4 was used; the minimum number of items to be presented to the S prior to any attempt to predict responses was set at one; no maximum difference between σ_M^2 and σ_B^2 was specified. A computer simulation was employed on the existing data, using a FORTRAN program written by the author and using the CDC 6500 computer at Florida State University, to predict the correctness of the item responses for Ss in the validation group.

Results

Descriptive statistics for the numbers of item responses obtained, correctly predicted, and incorrectly predicted per S, appear in Table 1.

TABLE 1

DESCRIPTIVE STATISTICS FOR ITEM RESPONSES

	Statistic		
Responses	Mean	Standard Deviation	Range
Obtained	8.37	2.75	[5,15]
Correctly Predicted	9.68	2.81	[4,14]
Incorrectly Predicted	1.94	1.54	[0, 6]

A phi coefficient was computed between predicted and actual responses on items for which predictions were made for Ss in the validation group. A value of .63 was obtained. For the marginal totals involved in the computation of phi, the maximum possible value (Guilford, 1965) was .91. A one-tailed t-test was performed to test the null hypothesis c = 0, in relation to the alternative hypothesis c = .63, where c represents the population correlation (Cohen, 1969). The null hypothesis was rejected ($t = 4.91$, $df = 36$, $\alpha = .01$, $\beta = .04$).

Descriptive statistics for the numbers of predictions per item and incorrect predictions per item appear in Table 2. The item statistics exclude consideration of the one item to which a response was obtained in each case prior to predictions. The ratio of the number of incorrect predictions per item to the total number of predictions for the item ranged from 0.00 to 0.48 with a mean ratio of 0.17.

TABLE 2

DESCRIPTIVE STATISTICS FOR
PREDICTIONS PER ITEM

	Statistic		
Predictions	Mean	Standard Deviation	Range
Total	23.26	7.24	[10,34]
Incorrect	3.89	3.07	[0, 11]

DISCUSSION

The tailored testing model proposed in this paper appears to be feasible. A data base containing the responses of more Ss might increase the effectiveness of the procedure for the following reasons: (1) the sampling variance of the difficulties of the items would be reduced; (2) the responses of outliers or Ss with unusual item response patterns would more generally be included; and (3) the number of observations constituting the denominator of a probability ratio would not as readily be less than the minimum number of required observations needed for prediction.

Additional research, using varying critical conditional difficulties, expected loss values, maximum differences in σ^2_M and σ^2_B, and numbers of items to be presented prior to any attempt at prediction might demonstrate more effective prediction than shown in this validation study. Analyses of the characteristics of items for which predictions are generally correct and items for which predictions are often incorrect might also be considered.

A purpose of this study was to attempt to predict the correctness of item responses. Further research might consider applying the proposed tailored testing model for the purpose of predicting mastery of objectives rather than individual items. A test consisting of objective-referenced items, with more than one item per objective, might be employed with the proposed testing model. The obtained and predicted responses for the items corresponding to the same objective would then be analyzed to determine if criterion performance might be predicted for the objective.

REFERENCES

Cleary, T., Linn, R., & Rock, D. An exploration study of programmed tests. *Educational and Psychological Measurement*, 1968, *28*, 345-360.

Cleary, T., Linn, R., & Rock, D. Reproduction of total test score through the use of sequential programmed tests. *Journal of Educational Measurement*, 1968, *5*, 183-187. (a)

Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press, 1969.

Cronbach, L. & Gleser, G. *Psychological Tests and Personnel Decisions*. Urbana, Illinois: University of Illinois Press, 1965.

Ferguson, R. A model for computer-assisted criterion-referenced measurement. *Education*, *91*, 25-31, Summer 1970.

Glaser, R. & Nitko, A. Measurement in learning and instruction. In Robert L. Thorndike (Ed.) *Educational Measurement*. Washington, D. C.: American Council on Education, 1971.

Green, B. Comments on tailored testing. In Wayne H. Holtzman (Ed.) *Computer-Assisted Instruction, Testing, and Guidance*. New York: Harper & Row, Publishers, 1970.

Guilford, J. *Fundamental Statistics in Psychology and Education*. New York: McGraw-Hill Book Co., 1965.

Hays, W. & Winkler, R. *Statistics: Probability, Inference, and Decision, Volume I*. New York: Holt, Rinehart & Winston, Inc., 1970.

Hubbard, J. Programmed testing in the examinations of the National Board of Examiners. (Collected writings) In A. Anastasi (Ed.) *Testing Problems in Perspective*. Washington, D. C.: American Council on Education, 1966.

Linn, R. Rock, D., & Cleary, T. The development and evaluation of several programmed testing methods. *Educational and Psychological Measurement*, 1969, *29*, 129-146.

Lord, F. Some test theory for tailored testing. In Wayne H. Holtzman (Ed.) *Computer-Assisted Instruction, Testing, and Guidance*. New York: Harper & Row, Publishers, 1970.

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

Veldman, D. *Fortran Programming for the Behavioral Sciences*. New York: Holt, Rinehart & Winston, Inc., 1967.

Waters, C. & Bayroff, A. A comparison of computer-simulated conventional and branching tests. *Educational and Psychological Measurement*, 1971, *31*, 125-136.