An Empirical Evaluation of Implied Orders as a Basis for Tailored Testing

Norman Cliff
Robert Cudeck
Douglas McCormick
University of Southern California

The tailored testing system described in this paper differs in several respects from others proposed for the same purpose. It does not require thousands of subjects for pretesting as do the tailored testing approaches used by the Educational Testing Service or the Civil Service Commission. It does not require concern about the wisdom of determining item statistics on one population and applying them in another. And it is an appropriate method for integrating testing into a training system.

The basic principle is simple. It arises from considering dichotomous items as furnishing ordering relations between persons and items. If the relations are consistent with each other, considered as a whole they furnish a joint ordering of the persons and the items. It is well known that the logical properties of an order are such that if certain of the relations among elements are known, the remainder can be deduced by making use of the transitivity property which characterizes orders. The basis of this approach to computer-interactive testing has been described by Cliff (1975). The general idea is that even an incomplete matrix of responses by persons to items can be used to deduce some order relations between items, which is their relative difficulty. These order relations in turn can be used to predict what the individual's responses will be to items not yet answered; therefore, the necessity of asking those items would be removed.

Taking a joint order as a model for test items is equivalent to assuming that the data provide a Guttman scale, but test data are not Guttman scales. However, a joint order is only an *approximate* model for test items. The problem in tailored testing is, then, one of modifying the transitivity principle in order to make it work reasonably well in the presence of error. A rough statistical approach is used here. At any given time, there are a certain number of responses implying that item j is harder than item k and a certain number that imply the reverse. If one kind of response predominates over the other, then it is implied that one item is easier than the other. Similarly, the pattern of responses by an individual to a subset of test items may be such that some of the responses imply that he/she should answer a particular test item, as yet untaken, incorrectly. Correspondingly, other responses may imply that he/she should answer it correctly. If one number predominates over the other, then the implication is made correspondingly.

Procedure

Illustrative Examples

Table 1 provides an illustration of the way in which the procedure operates. The two columns on the left show the responses of 15 persons to two items, j and k. To determine which of two items is easier, n_{jk}, the number who answered j correctly and k incorrectly, is examined in comparison to n_{kj}, the number who answered j incorrectly and k correctly. In the data illustrated, Person 5 was the only one who answered j correctly and k incorrectly, whereas Persons 4, 6, 7, 8, 9, 10, and 11 answered in reverse. The Frequencies 1 and 7 (n_{kj} and n_{jk}, respectively) are shown at the bottom of the table. Use of a statistical decision rule, which is outlined below, would lead to the decision that item k was easier than item j. For each pair of items in a test, such a comparison is made by means of the decision rule. The results of the comparisons are recorded in what is called the *item dominance* matrix. In the matrix a "1" means that the row item is more difficult than the column item.

Table 1
Illustrative Basis of TAILOR Process

	Complete				Incomplete		
	Items		Dominant		Items		Dominant
Persons	j	k	Item	Persons	j	k	Item
1	1	1	--	1	1	1	--
2	1	1	--	2	1		--
3	1	1	--	3		1	--
4	0	1	j	4	0		--
5	1	0	k	5		0	--
6	0	1	j	6			--
7	0	1	j	7	0	1	j
8	0	1	j	8	0	1	j
9	0	1	j	9	0		--
10	0	1	j	10			--
11	0	1	j	11	0	1	j
12	0	0	--	12			--
13	0	0	--	13		0	--
14	0	0	--	14	0		--
15	0	0	--	15	0	0	--
	Dominance Frequencies j = 7 k = 1 p = 9/256, therefore $j>k$				Dominance Frequencies j = 3 k = 0 p = 1/8, therefore $j>k$		

The foregoing is applicable to a complete test. In an incomplete or tailored test some of the responses would be missing, as is shown in the two righthand columns of the table. The quantities n_{jk} and n_{kj} can still be counted, however, since Person 5 has now only one item, $n_{kj}=0$. Of the seven persons who answered j incorrectly and k correctly, there is now data on both items from only Persons 7, 8, and 11; so n_{jk} is three. The comparison of n_{jk} to n_{kj} could still lead to the conclusion that item j was more difficult than item k if this were all the information available, provided that the liberal rule were used.

Now consider Person 2, who has answered the more difficult item correctly and has not taken the easier one. It could be concluded that he/she would answer the easier item correctly also, and therefore it would not be administered. Similarly, Person 13 has answered the easier item incorrectly; it could thus be concluded that he/she would answer the more difficult one incorrectly also if he/she were to take it, and it would not be administered. Actually, in making decisions of this kind, what is done is similar to deciding which items are easier and which more difficult. Suppose person i has not yet taken item j. The number of more difficult items he/she has answered correctly would be compared to the number of easier items he/she has answered incorrectly. If, by the same decision rule used earlier, the latter of these were to preponderate over the other, he/she correspondingly would be implied to have also incorrectly answered item i. If the reverse were true, then he/she would be assumed to have answered it correctly. If neither were to preponderate, then no decision would be made. At any given time, then, in the tailored testing process, as many inferences as possible are made about the relative difficulty of the items. These in turn are used to imply responses for each person to items he/she has not yet taken.

Frequency Comparison

A rather liberal two-part decision rule is used in comparing frequencies. The major part corresponds to comparing the frequencies by McNemar's (1969) binomial probability and rejecting the null hypothesis with a one-tailed alpha level of .33. Values of n_{jk} and n_{kj} of 2 to 0 and 3 to 1 thus lead to rejection, and an implication is made. The second aspect of the rule is used to deal with the instances where the frequencies are 1 and 0 only. If the information is very sparse (i.e., early in the testing process), even this small preponderance is used to imply item dominance or an implied response. (This is done by means of a complex probability evaluation which will not be detailed here.) The decisions are thus made on very small frequencies. Although until the end of the testing process a possibility exists that any one of them can be reversed, for the most part they remain quite stable.

In a sense this is not unique; any tailored testing system could fit the above description. What is unusual is the very small frequencies (as few as one) used to make the decisions and the simple decision rule employed. Perhaps the most unusual feature is that the process starts with no knowledge about the items; information on item difficulty is gained at the same rate as knowledge of the individual's abilities is obtained.

Modes of Operation

Group testing. There are two basic modes of operation, which might be termed simultaneous and cumulative. The simultaneous mode, called TAILOR (Cudeck, Cliff, Reynolds, & McCormick, 1976; Cudeck, Cliff, & Kehoe, in press), was developed first. It assumes that a number of subjects are being tested simultaneously with a particular pool of items and that there is no knowledge concerning the items. In the initial round of item presentations, items and persons are randomly assigned to each other for the first pairing. In subsequent rounds, each person is assigned the item that is currently closest to him/her in the joint partial order. The process ends when there is either an actual or an implied response for each person to each item.

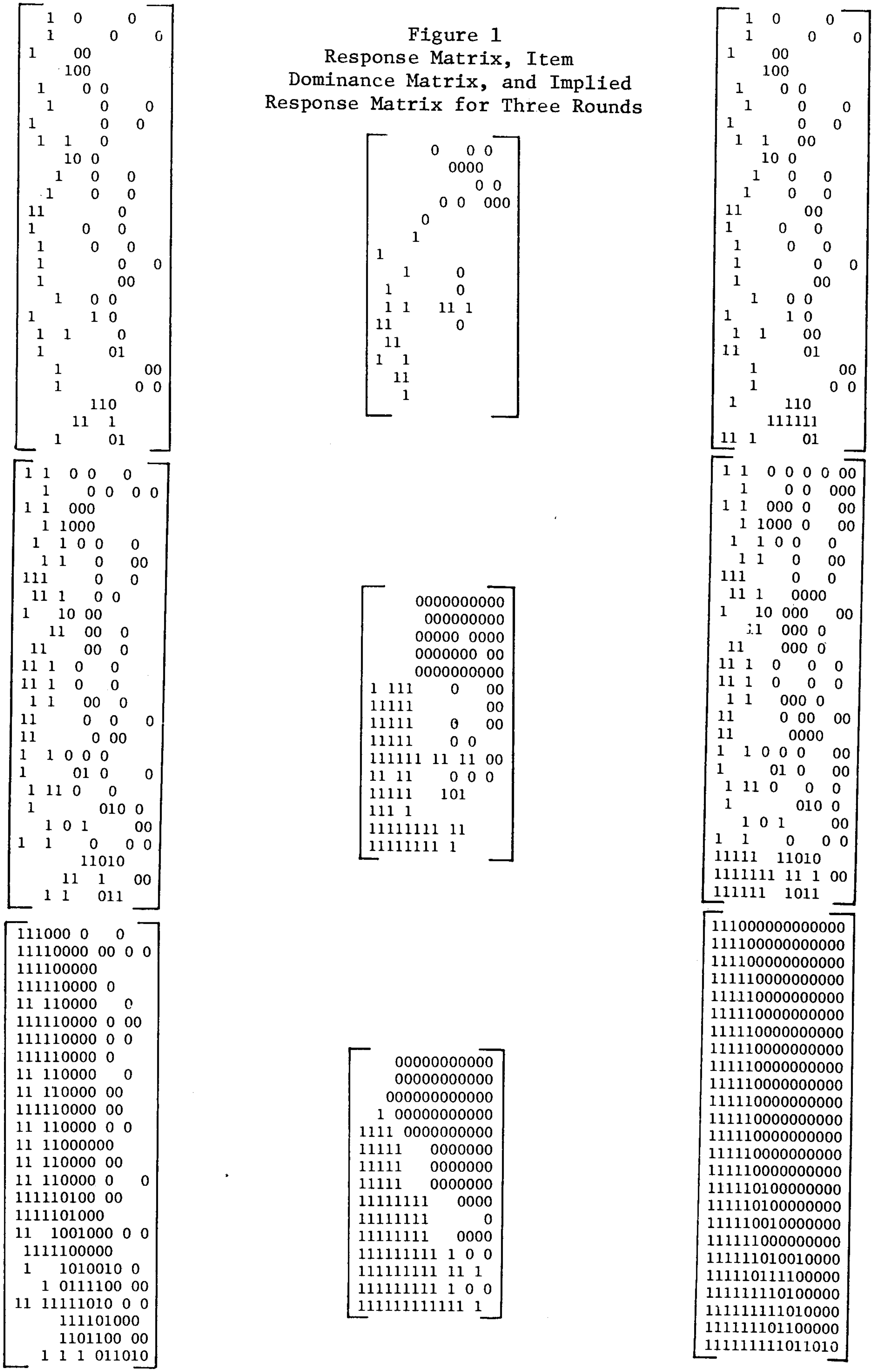
The means of deciding which item to assign to each person is the second major procedure of the process. For its complete operation, this approach carries out one further frequency-comparing step. Each person's implied response vector is compared to every other person's in order to compute a person-person dominance matrix by a means parallel to that used to obtain the item dominance matrix. That is, if person i is implied to answer more of the same items correctly which person h answers incorrectly than items which person h answers correctly and person i answers incorrectly, then i dominates (ranks above) h .

It is possible to assign a current total score to each item and person. For a person, this is the total number of items answered correctly (directly or by implication) minus the number answered incorrectly in the same way, plus the difference in the number of persons he/she dominates and is dominated by. For an item, this score is the number of persons who answer it incorrectly (directly or indirectly) minus the number who answer it correctly, plus the difference in the number of items it dominates and the number which dominate it. In this way items and persons are placed on the same ordinal scale, and the person takes the item for which he/she has no implied or direct response and which is closest to him/her on the scale. Given the various binary matrices involved, this process is actually very simple.

This mode of operation takes place by what might be called "rounds." At each "round," each person should be presented with an item. The item given at one round depends on the results of the previous rounds for that item and person, and each person participates in as many rounds as are necessary to complete his/her score vector.

This procedure is illustrated in Figure 1; the upper three matrices show the operation at any early stage of the process. The data are for 25 persons and 15 items on the Stanford-Binet. The matrices in the left column are the actual response matrices; the middle ones are the item dominance relations that are implied by them; and the right-hand matrices are the implied response matrices. In each a "1" means correct or dominance, a "0" means incorrect or antidominance, and a blank means no relation. The middle set of matrices

Figure 1
Response Matrix, Item Dominance Matrix, and Implied Response Matrix for Three Rounds



shows the operation at an intermediate stage of the testing process, and the bottom set shows the final stage; the matrix in the lower right-hand corner shows that the score matrix is now complete by implication.

Individual testing. The second mode of operation is sequential or cumulative and tests individual subjects only. This is called TAILOR-APL (McCormick & Cliff, in press). Again, no knowledge about the items is assumed. The first person must, therefore, take all items. After a few persons have taken the items, however, there may be enough information to define the relative difficulty of some items. This information is then used to infer the responses of subsequent persons to these items, thereby removing the necessity of taking them. As more and more information accumulates, more relative difficulty relations also accumulate, so that the tests become more and more "tailored" for later subjects.

Data

There are now three kinds of data on one or the other of these methods: (1) monte carlo studies assuming a stochastic model; (2) real-data simulations using data files from complete tests; and (3) actual live tailored testing.

The main dependent variable in each is either the correlation of obtained scores with true scores or correlations between the scores on parallel forms. The major comparisons are with these variables under tailored and complete testing conditions. Of additional interest are a number of variables reflecting cost and efficiency factors and the effects of such elements as statistical parameters of the item pool and circumstances of testing on the results.

Monte Carlo Study

The most extensive data comes from a monte carlo study based on applications of Birnbaum's (1968) three-parameter model. A variety of different characteristics for an item pool were assumed, and a certain number of items were sampled from hypothetical pools with the prescribed characteristics. These item pool characteristics included the mean and standard deviation of the item discrimination parameter, the mean and standard deviation of the item difficulty parameter, and the mean and standard deviation of the item guessing probability. Difficulty and discrimination were assumed to be normally distributed with the prescribed mean and variance; sometimes the variance was zero. It was assumed that a certain number of subjects were being tested simultaneously, and their true scores were sampled from a standard normal population. The program was put in operation and a random number generator was used in conjunction with the Birnbaum model to determine the correctness of each response.

A variety of different combinations of conditions were run; they are given in Table 2. This paper is primarily concerned with a particular subset of conditions. This was a 2 × 2 × 2 factorial design where the variables were (1) number of subjects, 25 or 40; (2) number of items, 15 or 25; and (3) mean discrimination index, a=1.0 or 2.0. Mean item difficulty was zero, as was the variance of item discriminations and the value of the guessing parameter. It should be noted that sampling fluctuations can lead to appreciable

Table 2
Characteristics of Samples of Score Matrices
Generated by Latent Trait Models

		Item Discrimination		Item Difficulty		Mean
Persons	Items	Mean	*S.D.*	Mean	*S.D.*	Guessing
10	25	1	0	0	1	0
10	25	2	0	0	1	0
25	15	.5	0	0	1	0
25	15	.5	0	0	1	0
25	15	1	0	0	1	0
25	15	2	0	0	1	0
25	15	1	0	1	1	0
25	15	2	0	1	1	0
25	15	1	0	0	2	0
25	15	2	0	0	2	0
25	15	1	.2	0	1	0
25	15	2	.2	0	1	0
25	15	2	.4	0	1	0
25	25	1	0	0	1	0
25	25	2	0	0	1	0
25	25	1	0	0	1	.1
25	25	2	0	0	1	.1
25	25	1	0	0	1	.2
25	25	2	0	0	1	.2
25	25	1	.2	0	1	0
25	25	2	.2	0	1	0
25	25	2	.4	0	1	0
40	15	1	0	0	1	0
40	15	2	0	0	1	0
40	15	1	0	0	1	.2
40	15	2	0	0	1	.2
40	25	1	0	0	1	0
40	25	2	0	0	1	0

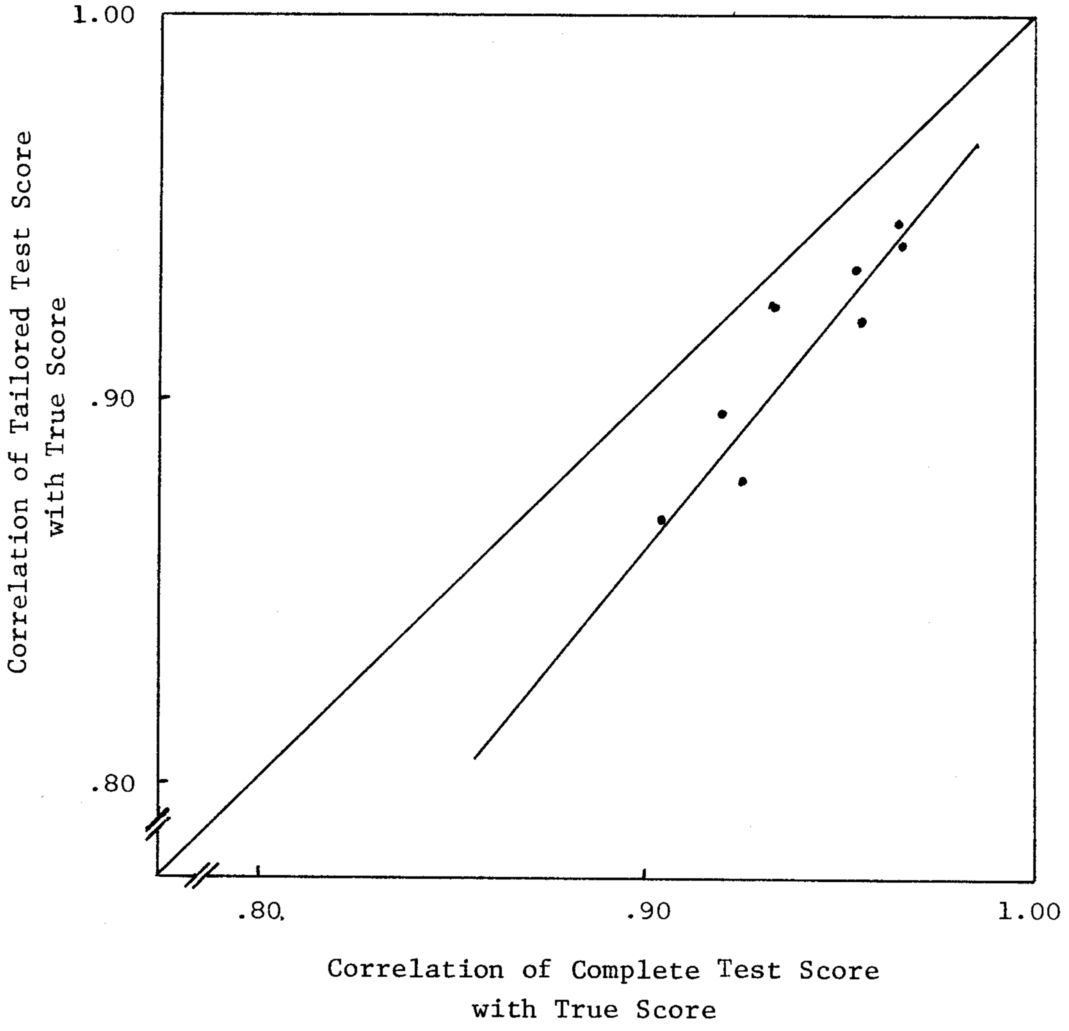
mismatch in difficulty. The average "person" received about half the items under these conditions. More items and/or more persons meant a smaller fraction of items per person, on the average, as shown in Table 3.

Table 3
Proportion of Items Used in Monte Carlo Data

	25 Persons				40 Persons		
Item	Items			Item	Items		
Discrimination	15	25	Mean	Discrimination	15	25	Mean
1.0	.609	.514	.562	1.0	.543	.471	.507
2.0	.580	.516	.548	2.0	.548	.441	.494
Mean	.594	.515	.555	Mean	.546	.456	.501

Of more interest, perhaps, is the comparison of the correlation with true score for complete and tailored tests. For these data, the validities were .940 and .913, respectively. Thus, the tailored validities based on half the items were, on the average, 97% as high as those for the complete data. Figure 2 provides more detail, showing a rather close relation between the validity of the complete test and that of the tailored test. Each of the eight points corresponds to the average of five replications of one of the eight combinations of item discrimination parameters, number of items, and number of persons in the 2 × 2 × 2 factorial design. The regression line in Figure 3 may be compared to the 45 degree line which is also indicated. As expected, validity of the tailored test was below that of the complete test; but there was a close correlation. Tailored validity, however, appeared to fall off more steeply than complete validity, i.e., the slope was greater than 1.0.

Figure 2
Relation Between Complete and Tailored Test Validities



Data for all the combinations of conditions showed essentially the same picture: By far the major determinant of tailored validity was complete test validity. Anything that affected the latter (i.e., mean discrimination, guessing probability, number of items in the pool) also affected tailored test validity. Furthermore, the effect was clearly somewhat disproportionate; reducing complete test validity reduced tailored validity even more.

Item File Data

A file of the responses of 625 children from ages 2 to 15 to the Stanford-Binet items was used to simulate the process of actual testing using the simultaneous procedure. The children were divided into three subgroups on the basis of chronological age. Within each subgroup, two samples of 25 items were drawn. A total score was computed on each set of items. One set was used in the TAILOR procedure. The major outcomes of interest were the number of items used in the latter and the correlations of the resulting score with the total score on the untailored half. The simulations were done with either 20 or 40 persons assumed to be tested simultaneously. Within each age group, five samples of 20 and five of 40 were drawn.

The results were quite similar to those for the monte carlo data. About half the items were presented in each case, 55% when 20 persons were tested and 46% when 40 persons were tested. The correlation of the tailored test scores with the complete half-test scores averaged .85 (see Table 4), whereas the correlation between scores on the two complete halves was .88. Thus, the ratio of complete to incomplete correlations was coincidentally again .97. Neither among the age-groups nor between group-sizes were there significant variations. That the latter correlations were somewhat higher in Table 4 is apparently a sampling accident. Of further interest is that responses to 96% of the items not taken were correctly predicted by the procedure.

Table 4
Average Correlations of Tailored and Complete Tests with a Complete Parallel Form in Binet Data

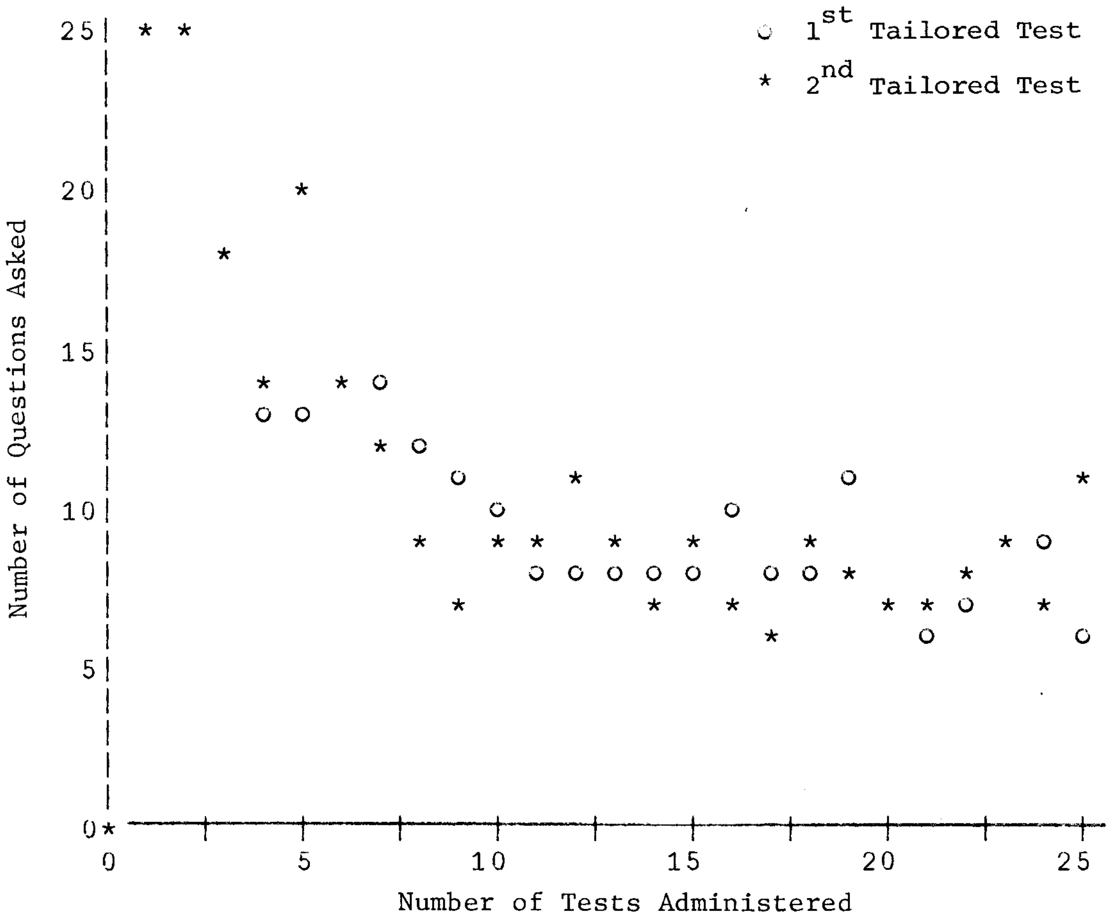
Persons	Comp-Comp	Comp-Tailor
20	.855	.829
40	.889	.866

Live Testing Data

To date, the data with human examinees is only available from the cumulative testing mode, TAILOR-APL. With that procedure only 25 subjects have been used in the tailored and complete testing conditions. The test used was composed of anagrams (scrambled words); it was easy to write a scoring routine for it in a non-multiple-choice mode. It was felt that avoiding a multiple-choice format was desirable on the basis of the results of the monte carlo studies. The computer typed a scrambled word, and the subject's task was to type back the correctly unscrambled word within a specified time limit.

In this mode, too, the items were divided into halves of 25 items each. This time, however, a subject either took two tailored tests simultaneously or two complete tests, rather than one of each as simulated with the Binet data. This procedure may have been important.

Figure 3
Relationship Between Number of Questions Asked and
Number of Tests Administered for First and Second Tailored Tests



It should be remembered that in the cumulative mode, the first subject takes all the items, and each subsequent subject takes progressively fewer items. In Figure 3 the data for the tailored subjects seems to be headed toward an asymptote of around 8 items, even though the average is 11. This given an idea of the rapidity with which this can take place. A different view of the process is presented in Figure 4. Here the box means "correct," the star means "incorrect," and blank means "not taken."

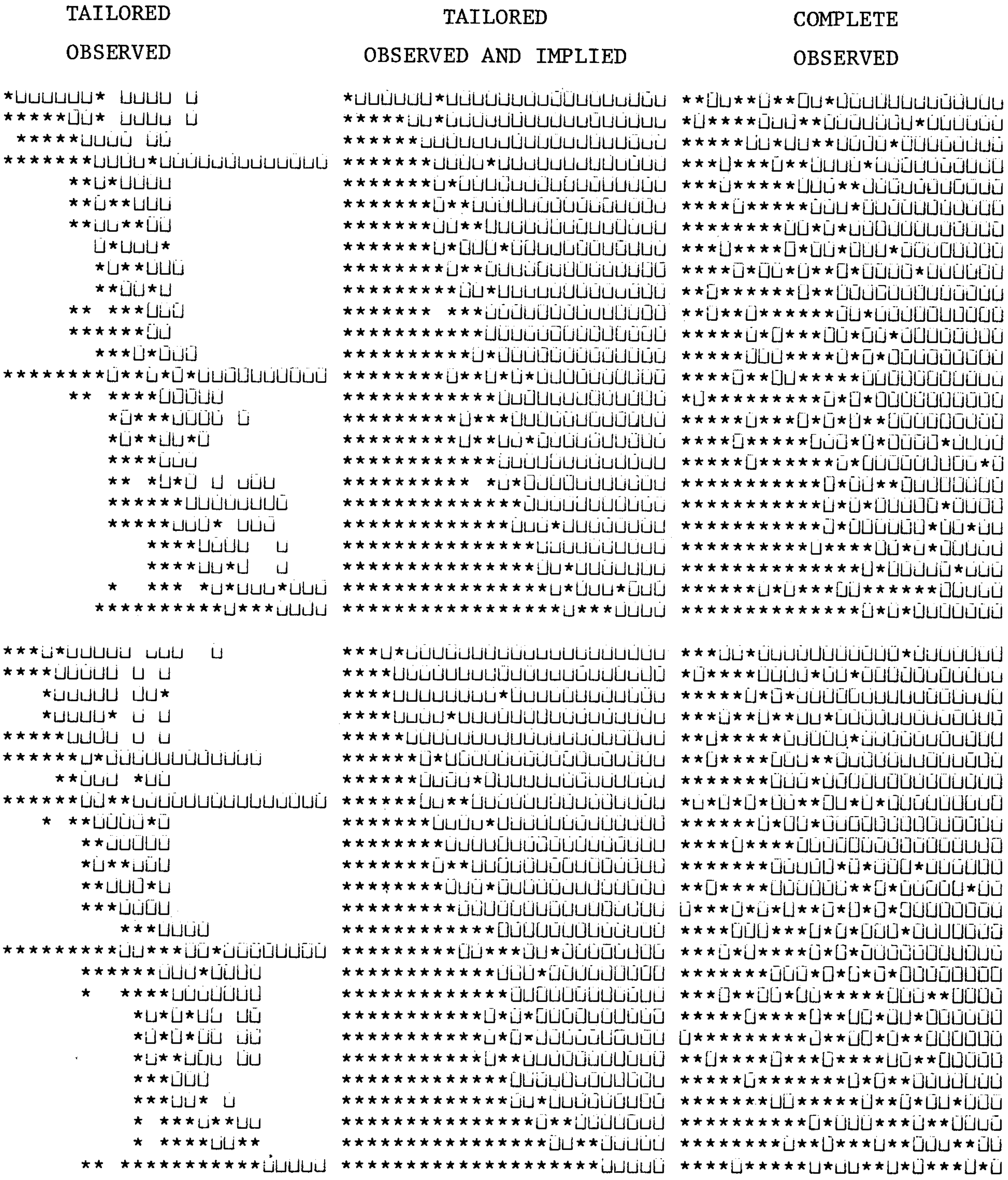
Figure 4
Observed Responses in Each Tailored
Test Arranged Chronologically

```
*******□□□□*□□□□□□□□□□□□□ ******□□**□□□□□□□□□□□□□□□
********□**□*□*□□□□□□□□□□ *********□□***□□*□□□□□□□□
       **********□***□□□□       ** ***********□□□□□
*□□□□□□* □□□□ □           ***□*□□□□□ □□□  □
*****□□* □□□□ □           ******□*□□□□□□□□□□□□
        *  *** *□*□□□*□□□       ******□□□*□□□□
        ******□□□□□□□□          *  ****□□□□□□□
        *****□□□* □□□                 * ****□□**
     ** ****□□□□□                   *□**□□□
        *□***□□□□ □                     *□*□*□□ □□
     ** ***□□□                 * **□□□□*□
     **□**□□□             ****□□□□□ □ □
        *□**□□*□                        *□*□*□□ □□
     ******□□                   **□□□*□
     **□□**□□                           *□**□□□ □□
        ** *□*□ □ □□□           ***□□□□
       ***□*□□□                         ***□□□
     **□*□□□□                *□□□□□ □□*
 *****□□□□ □□                *□□□□* □ □
        ****□□□                         ***□□* □
       **□□*□                   **□□□□□
       *□**□□□                *□□□ *□□
           ****□□□□  □                  * ***□**□□
           ****□□*□  □                 ***□□□□
       □*□□□*             *****□□□□ □ □
```

The subjects are arranged from earliest to latest, and the two panels show the results for the two subtests. The items in each are ordered from left to right in terms of the final order of difficulty.

Even though an average of half the items were presented, the reliability of the tailored test was substantially, although not significantly, higher than for the complete test (.81 vs. .66 with N_1=N_2=25). This is shown quite strikingly in the response matrices for the two conditions, as depicted in Figure 5, where the items are ordered in terms of the difficulty and the persons in terms of scores. The left panel is the incomplete data, the middle panel is the complete matrix inferred from it for the same persons, and the right panel is the corresponding score matrix for the complete data. The two 25-item pools are shown in the upper and lower halves. What appears striking is the substantially greater regularity apparent in the tailored matrix as compared to the complete one. It is much rarer for a person to answer an easy item incorrectly and a difficult one correctly, or vice versa. It appears that the subjects were behaving more consistently under the tailored condition, even though a statistical demonstration of this cannot yet be made.

Figure 5
Response Matrices



Discussion

If the appropriate variation on the Spearman-Brown formula is used to compare tailored and complete reliabilities or validities for the simulations, and the formula is solved for the length factor, it appears that the tailored test behaves like a complete test in the simulations with about 25% more responses. This modest saving may be the best that can be done in any system *if pretesting to determine item parameters is included.* Indeed, even from an information-function point of view, it would seem reasonable to conclude that the approach to tailored testing presented here may seem the most plausible. It makes use of the information-function for both items and persons simultaneously, albeit in an informal manner. That is, administering a difficult item to a low ability person does not give much information about either the item *or* the person. Therefore, from the beginning there is an attempt to match items and persons appropriately and thereby obtain more information about both.

Designs Used in Evaluating Tailored Testing

Some type of cross-validation approach is necessary for a tailoring method that relies on prior estimates of item parameters if a reasonable estimate of its efficacy is to be derived. This must be done in a realistic way; spurious correlations between non-independent scores must be avoided. With real data it is at least necessary to estimate parameters on one sample and use them to tailor the test for a second sample. These could be samples from the same population, but it is more realistic to do such things as estimate parameters on data from one year and use it to tailor in a second year. In these ways, the effects of sampling fluctuations in the estimates or item parameters can be more realistically mirrored, since this is how the system would operate in practice.

In the second sample, scores on parallel forms must be available. There must be either two separately tailored and administered tests or a tailored test and a separate conventional test. The parallel form correlation between them must be calculated and then compared to two conventional parallel forms. In this way spurious estimates of agreement are avoided.

If the study is of a monte carlo nature, two samples are still necessary: one for estimating item statistics for use in tailoring and the other to apply them in a tailored test. At present, parallel forms do not seem necessary, since correlations can be presumably calculated with true score and compared with the corresponding correlation for a conventional test. This will give a basis for comparison.

The necessity of pretesting should also be taken into consideration in evaluating efficiency. If 1,000 people take 100 items each in order that a second 1,000 can be given a tailored test of only 20 items, then it seems that the savings due to tailoring are only 40% rather than 80%. Tailoring is only effective if the sample on which item statistics are determined need be only a fraction of that to which the tailored form will be administered. These three considerations--independent samples, independent measures, and inclusion of pretesting costs--seem necessary to be included in the assessment of the usual tailored testing. Of the three, only the necessity for the availability of independent scores seems necessary for the assessment of the approach presented here, however, since the single administration strategy obviates the need for a separate norming sample.

The third point is that the computer may bring psychology back into testing. The subjects seemed to *behave* differently under the tailored condition, perhaps their minds even worked differently. If the data are taken at face value, the eleven tailored responses acted like a 55-item complete test. It appears that the subjects simply behaved more consistently on the tailored test; a high ability subject was less likely to give an incorrect answer to an easy item and a low ability subject was less likely to correctly answer a difficult item.

Cost

The monte carlo studies cost about an average of ten cents per pseudosubject; the efficiency of the program can probably still be increased by a factor of two to five, and computer costs will reduce by a similar factor in the next three or four years. Therefore, this aspect of the cost of testing will be relatively small in most applications compared with, for example, item writing. The actual computer costs of administering the anagrams test to one person was about $3.50; current revisions of the program should reduce this to below $1.00.

Conclusions

The tailored testing procedure described in this paper appears to be cost-effective, applicable to small populations, and relatively free from the encumbrances of an item traceline model. Thus, it might be a viable alternative to other approaches to tailored or adaptive testing.

References

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

Cliff, N. Complete orders from incomplete data: Interactive ordering and tailored testing. Psychological Bulletin, 1975, 82, 289-302.

Cudeck, R., Cliff, N., & Kehoe, J. TAILOR: A FORTRAN program for interactive tailored testing. Educational and Psychological Measurement, in press.

Cudeck, R., Cliff, N., Reynolds, T., & McCormick, D. Monte carlo results from a computer program for tailored testing (Technical Report No. 2). University of Southern California, Department of Psychology, 1976.

McCormick, D., & Cliff, N. TAILOR-APL: An interactive program for individual tailored testing. Educational and Psychological Measurement, in press.

McNemar, Q. Psychological statistics (4th ed.). New York: Wiley, 1969.

Acknowledgements

This research was supported by the Office of Naval Research, Contract N00014-75-C-0684, NR 150-373. Appreciation is extended to Mark Reckase of the University of Missouri for making available the item file data from the Stanford-Binet. The live testing data are from a Master's Dissertation written by Douglas McCormick.