Monte Carlo Evaluation of Implied Orders As a Basis for Tailored Testing

Robert Cudeck, Douglas McCormick, and Norman Cliff
University of Southern California

TAILOR, the computer program which implements an approach to tailored testing outlined by Cliff (1975), was examined with errorless data by monte carlo methods. Three replications of each cell of a 3 × 3 table with 10, 20, and 40 items and persons were analyzed. Mean rank correlation coefficients between the true person and item order, specified by preassigned random numbers, and the computed order produced by the program averaged .96. The average proportion of items used was .48. A marked tendency was observed for the program to produce better results as the number of persons and items increased.

This paper reports an evaluation of a system for tailored testing proposed by Cliff (1975). The system makes use of the fact that responses to test items contain two kinds of information—one which is *explicit,* based on observed person-item relations, and one which is *implicit*, pertaining to inferred relations among persons or among items. The model for relating these sources of information is the Guttman scale, but since test data do not conform to perfect scales, the method contains a procedure for handling inconsistent responses. The present article outlines the general aspects of the procedure, which is called TAILOR, and then describes the design and results of a monte carlo study which attempts to assess the performance of the method with errorless data.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 3, No. 1 Winter 1979 pp. 65-74

Implied Orders Tailored Testing

In the following discussion, it is convenient to define a basic supermatrix which is composed of four sections. First, let S be the usual response matrix which is persons-by-items and which contains correct responses, i.e., $s_{ij} = 1$ if person i passes item j, and zero otherwise. Corresponding to S, $\tilde{S}$ records the incorrect responses, such that $\tilde{s}_{ij} = 1$ only if person i misses item j. Normally, S and $\tilde{S}$ are direct counterparts of each other, but with tailored tests this will not usually be the case. Therefore, both matrices are required in order to keep track of responses which are correct, incorrect, or not yet asked. An item dominance matrix is computed in $N = \tilde{S}'S$, with elements n_{jk} equal to the number of persons who failed item j and passed k. The corresponding person dominances are contained in $X = S\tilde{S}'$, and x_{ih} equals the number of items that person i passed and person h failed. The explicit information is contained in S and S′; the implied data is in N and X. The supermatrix A thus is items-by-persons by items-by-persons as is shown in Figure 1.

The testing process itself is actually an iterative procedure which begins by asking each person an item of median difficulty. Since no pre-

Figure 1
Full supermatrix with incomplete data

$$A = \begin{bmatrix} N & \tilde{S}' \\ S & X \end{bmatrix} =$$

	Items				Persons				
Items	0	0	0	0	0	1	1	1	0
	0	0	0	0	0	0	1	1	0
	0	0	0	0	0	0	0	1	1
	0	0	0	0	0	0	0	0	1
Persons	1	1	1	0	0	0	0	0	0
	0	1	0	1	0	0	0	0	0
	0	0	1	1	0	0	0	0	0
	0	0	0	1	0	0	0	0	0
	0	0	0	0	0	0	0	0	0

testing is required, during the first phase of item assignment the pairs are made at random, and the submatrices N and X are null. Item and person dominances result from powering A, $A^2 = AA$, as shown in Figure 2. In A^2, N and X are integer in form. Also, if the responses are consistent, they will display an upper triangular pattern which is characteristic of the Guttman scale. Cliff (1975) describes how the information in X and N can then be used to infer an order of ability and difficulty, respectively. The primary characteristics of orders are asymmetry and transitivity, properties which are apparent in Figure 2. Thus, the next step in the process is to transform the integer dominance relations to binary relations which show the implied order between persons or between items.

Note that this process is consistent with the ordinal nature of the testing theory being used. The observed data in S and $\tilde{S}$ are indicative of binary orders which characterize dominance of a person over an item or an item over a person. To extend the idea of ordinal relations to item or person dominances requires that a method be given for inferring the binary relationship of simple dominance from integer entries in N and X. This requirement means that one *imply*, based on the pattern of dominances, that a gen-

Figure 2
Integer item and person dominance matrices

$$A^2 =$$

	Items				Persons				
Items	0	1	1	2	0	0	0	0	0
	0	0	1	2	0	0	0	0	0
	0	0	0	1	0	0	0	0	0
	0	0	0	0	0	0	0	0	0
Persons	0	0	0	0	0	1	2	3	1
	0	0	0	0	0	0	1	1	1
	0	0	0	0	0	0	0	1	2
	0	0	0	0	0	0	0	0	1
	0	0	0	0	0	0	0	0	0

uine order exists between members of the set. For example, given the elements n_{jk} and n_{kj} from N, determine the ratio

$$z_{jk} = \frac{n_{jk} - n_{kj}}{(n_{jk} + n_{kj})^{\frac{1}{2}}} \qquad [1]$$

where Equation 1 is the ratio of correlated proportions (McNemar, 1947). Then by using a prespecified criterion for z_{jk}, n_{jk} is set to 1 when the obtained ratio exceeds the criterion; n_{kj} is recorded as 1 if $-z_{jk}$ exceeds it; and both are zero if neither is the case. Thus, j dominates k only if answered incorrectly by "significantly" more persons. Some other functions could obviously be used for this purpose, but this is the approach used here. The same procedure is used for person dominance relations. It is important to note that this procedure modifies N and X in the supermatrix A^2 from integer to binary entries.

Next, suppose there is an item which person i passes but h fails, and a second item which is in turn failed by person i. In this simple illustration, it is implied that i dominates h. Since i fails the second item, and since h has less ability than i, it must be the case that h would fail it too. Similarly, if there is an item which h passes, it need not be presented to i because he/she must pass it. Computationally, A^2A gives a matrix in which submatrices N and X are again null, while S and $\tilde{S}$ are integer in form. The entries in S are the number of times person i actually dominates item j, plus the number of times he/she is implied to dominate j. In $\tilde{S}$ the entries record the actual and implied dominances of item j over person i; see Figure 3.

The integer entries in A^2A are then modified in a manner similar to Equation 1 to put the relations into binary form. However, since the test is between members of different sets, Equation 1 becomes

$$z_{ij} = \frac{s_{ij} - \tilde{s}_{ij}}{(s_{ij} + \tilde{s}_{ij})^{\frac{1}{2}}} \qquad [2]$$

In this instance, s_{ij} and $\tilde{s}_{ij}$ are integer values from A^2A. Again, if z_{ij} is greater than the criterion, then person i dominates item j; if $-z_{ij}$ is greater than the criterion, j dominates i. In effect, new submatrices S and $\tilde{S}$ are constructed which contain binary entries based on the integer values of A^2A.

The final result of this procedure is simply to use the rules of Boolean addition to add together the binary forms of $A + A^2A$. This last supermatrix contains all the relations between the

Figure 3
Integer matrices of person-item implications

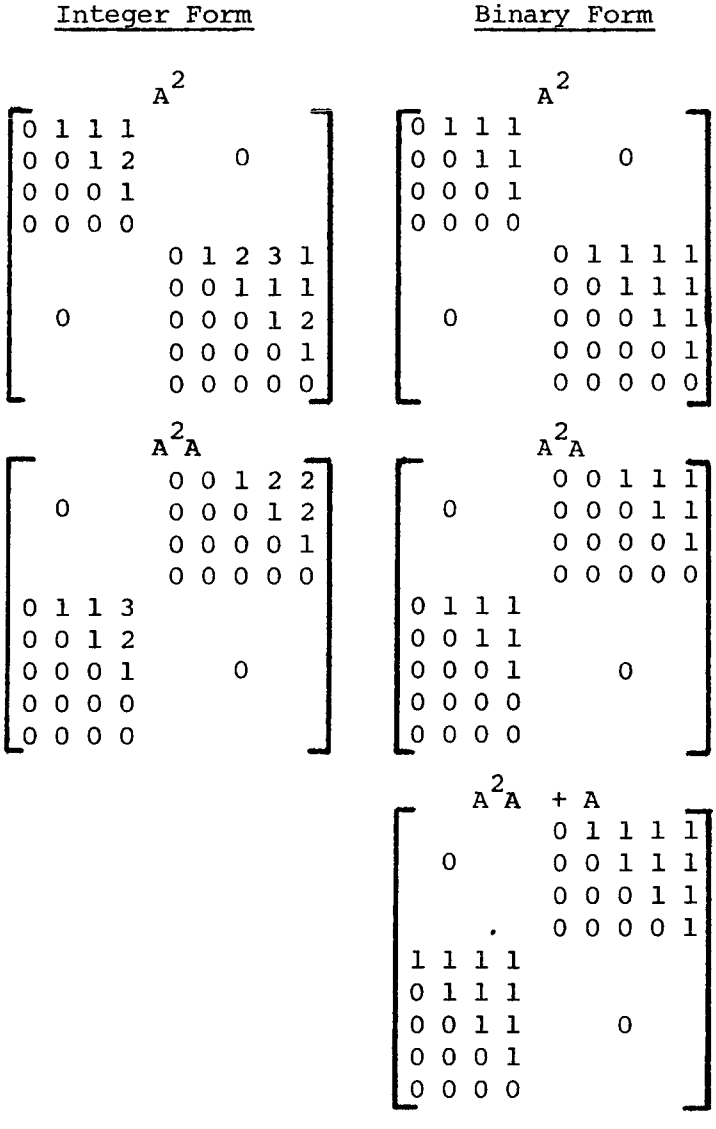
$$A^2A = \begin{array}{c} \\ \text{Items} \\ \\ \\ \\ \\ \text{Persons} \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{cccccccccc} \multicolumn{4}{c}{\text{Items}} & & \multicolumn{5}{c}{\text{Persons}} \end{array} \\ \left[\begin{array}{cccccccccc} 0 & 0 & 0 & 0 & & 0 & 0 & 1 & 2 & 3 \\ 0 & 0 & 0 & 0 & & 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & & 0 & 0 & 0 & 0 & 0 \\ & & & & & & & & & \\ 0 & 1 & 1 & 3 & & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2 & & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & & 0 & 0 & 0 & 0 & 0 \end{array}\right] \end{array}$$

items and persons which are actually observed or implied from the results of Equations 1 and 2.

Figure 4 shows the entire procedure, which is summarized below.

1. Integer dominance matrices N and X are computed in A^2 and dichotomized according to Equation 1 in the binary version of A^2.
2. Integer person-item matrices S and $\tilde{S}$ are given by A^2A and yield binary relations by means of Equation 2 in the binary version of A^2A.
3. By Boolean addition, compute $A + A^2A$.

Figure 4
Three-step procedure for obtaining the summary response matrix of obtained and implied relations



The beginning of a second round in this iterative process again determines item-person pairs. When no information is available, this is done by random assignment. After even one round, however, the scores from the initial questions can be used to select an item of approximately appropriate difficulty for each person. The traditional score for a person is his/her sum across a row of the score matrix S; the traditional item difficulty (actually, it is an easiness) is the sum down a column of the score matrix S, divided by n. The present formulation extends this method by considering all four components in the supermatrix, some of which contain observed relations and others which are implied. To bring together all aspects of the test, compute by Boolean addition

$$G = (A + A^2A) + A^2 \qquad [3]$$

where the term $A + A^2A$ is the binary form of S and $\tilde{S}$ and A^2 is the binary form of N and X. A net dominance score for each person on n items is the difference between the total wins and the total losses, where the total wins are the number of 1's in a given row of G and the losses are the number in the corresponding column. For the next round, a person is given the item with the net score nearest to his/her score. Testing continues until each person has an actual or implied relation with each item.

Before a testing system such as that proposed here can be put into actual use, a thorough examination of its performance under controlled conditions is desirable. The most straightforward design is one in which the persons and items are simulated by means of the monte carlo method. Essential information regarding the effects of sample size and decision criteria on speed and efficiency may then be obtained.

Method

This experiment used errorless data and assumed that no prior information was available for either persons or items. For each of n_p "persons" and n_i "items," a uniform random deviate x_k, $k = 1, n_p + n_i$, $0 \leq x_k \leq 1$, was generated from the method described by Knuth (1973, Vol. 2, chap. 3). These values served as measures of person ability or item difficulty such that whenever the number for person p_i was greater than that assigned to item i_j, then p_i was said to have answered i_j correctly. When the converse was true, then p_i was said to have missed i_j. No provision was made in this elementary model for chance success due to guessing or for the effects of item discrimination.

The major independent variable examined was the number of persons and items: n_p, n_i = 10, 20, 40. The dependent variables of interest were (1) the rank correlation between true score order and response score order; (2) the percentage of responses required for a complete solution; (3) the rate at which implications were made; and (4) the amount of computer central processing time (CPU) required. The decision criterion for the ratio for correlated proportions was set to 1.0. Each cell of the 3 × 3 design was replicated three times.

Results

Correlation of True and Estimated Rank Orders

The solution which TAILOR produces is a joint rank ordering of the persons and items along the ability-difficulty continuum. The major outcome criterion is the correlation of the computed rank order given at the end of a monte carlo run, and the true rank order of the persons and items based on the initially assigned random numbers. The mean rank order correlation coefficients based on Kendall's Tau b for the nine sample sizes are shown in Table 1. All the validities were close to unity. As can be seen, there was a tendency for studies with larger numbers of relations to produce somewhat higher correlations, presumably because the greater the number of possible pairs, the more likely the computation of a dominance relation between any two elements becomes (that is, be-

Table 1
Mean Correlation Coefficients Between True Order and Computed Order

Number of Persons	Number of Items		
	10	20	40
10	.94	.97	.93
20	.95	.98	.98
40	.95	.98	.99

tween any item-person, item-item or person-person).

It is noteworthy that with this errorless data, there were no perfect validities for any of the studies. This situation arises when the random assignment of true scores places two persons, for example, immediately adjacent to each other in the joint order, with no item intervening. In such cases, although the person or item pairs are not out of order, they are in fact not in perfect correspondence, so the correlation is less than unity. Consequently, these results suggest that TAILOR was highly successful in recovering the order of the original true scores. The fact that the validities were less than unity indicates that the order among any adjacent persons remains indeterminate when no item exists such that $p_i \geqslant i_j \geqslant p_q$.

Number of Responses

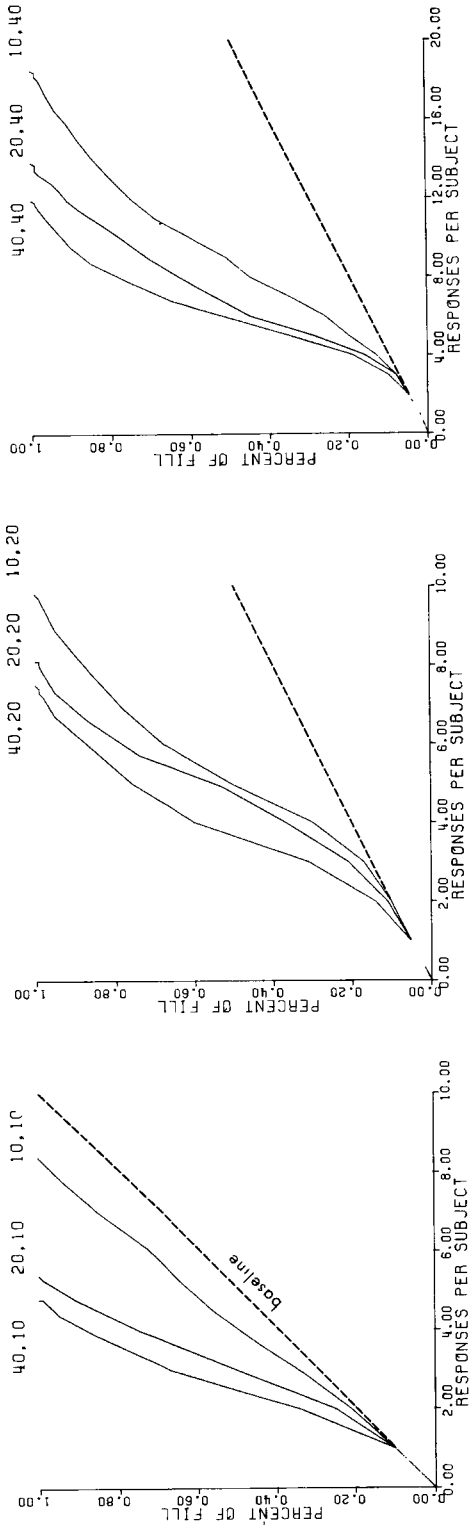
A second major outcome variable was the number of responses needed for a complete solution to be obtained. The ultimate value of a tailored testing approach lies in how much information a single response produces or, conversely, in how many relations are eliminated by the decision rules. Table 2 shows the mean percentage of responses which were required for each data set size. The values ranged from 34% to 72%. Again, there was a strong effect in the larger studies for sample size, such that as more persons or items were tested, relatively fewer responses were required.

The relation between the number of responses, the rate at which item-person implications were made, and the effect of sample size is given in Figure 5. The three panels separate the data by number of items. The abcissa contains the number of responses given by each person, while the ordinate shows the number of responses-plus-implications as a percentage of the product $n_p n_i$, which has been labeled the percentage of fill. If no implications were ever made, the total number of relations would always equal the number of responses, and the percentage of fill would be a perfect linear function of responses. This minimum performance

Table 2
Mean Percentage of Possible Relations Accounted for by Responses

Number of Persons	Number of Items		
	10	20	40
10	.72	.49	.46
20	.58	.44	.34
40	.48	.39	.46

Figure 5
Rate of implications as a function of the number of persons and items



level has been marked by the dashed line labeled base rate. When any implications are made, the percent of fill will depart from the base rate, such that the greater the difference in slope between base rate and percentage of fill, the more efficient is the tailoring of the test.

As can be seen, in the 10-item test relatively few implications were ever made; when this did occur, it did so at a very even, nearly linear rate. However, in the 20-item data sets, and especially in the 40-item ones, there was a marked departure from the base rate performance. The majority of the implications tended to be made after about three or four responses per person had been obtained and fell off when about 80 to 85% of the relations were accounted for. There was also a consistent, although somewhat less dramatic, effect for the number of persons.

CPU Time

Finally, performance was also assessed in terms of central processing unit (CPU) time. CPU time is the actual amount of time a computer system is involved with calculation or the institution of input and output. It is not invariant from machine to machine, but the following figures should provide a rough estimate of the relative expense of using TAILOR in other environments. In fact, shortly after this study was conducted, a series of programming modifications was undertaken to greatly reduce the time requirements on the IBM 370/158 available to the authors. These changes have resulted in CPU time reductions of 1/2 to 2/3. Consequently, the amounts of processing time reported below should not be taken as absolutes.

Table 3 shows the average amount of CPU time in seconds for the nine conditions, as well as the mean across items for n_p persons and across persons for n_i items. In each case, as $n_p n_i$ increases, so did CPU time. However, the number of items contributed to the total time at a much greater rate than the number of persons. This can clearly be seen in Figure 6, which contains CPU time as a function of the marginal means of Table 3. The testing of additional persons merely had an additive effect on CPU time, but increasing the number of items increased the computing requirements in an exponential fashion. This finding is not surprising, since TAILOR works from the item information rather than from the person information to compute the person-item implications. Designing the algorithm in this manner makes the number of items the primary factor affecting CPU time. The advantage of this method is that when a researcher increases the number of subjects, little is added to the amount of computer time needed.

Even though CPU time increased whenever $n_p n_i$ increases, the cost-per-subject rates were quite modest. For example, with $n_i = 40$ and $n_p = 10$, 20, or 40, the CPU time from Table 3 was 64, 89.4, and 132.1 seconds, respectively. This results in an average of 6.4, 4.5, and 3.3 seconds per subject, which at the standard USC computing rates of $4.00 per minute was .43, .30, and .22 cents per person. Although no figures are available on the comparable costs of conven-

Table 3
Mean CPU Time*

Number of Persons	Number of Items			Mean
	10	20	40	
10	3	10.8	64	25.9
20	5.2	18.7	89.4	37.8
40	11.9	36.6	132.1	60.2
Mean	6.7	22	95.2	

*Central Processing Unit time for IBM 370/158 with version 1.0 of TAILOR. Subsequent versions are substantially faster.

Figure 6
Effects of the number of persons or items on central processing (CPU) time

tional testing, these rates suggest that implied orders tailored testing may not be completely infeasible economically.

Discussion

This preliminary investigation of TAILOR involved errorless data and data sets ranging in size from 100 to 1,600 relations. The findings were generally positive. All outcome measures behaved in an orderly fashion across the various study sizes. It has been suggested by Knuth (1973, Vol. 3) that the minimum number of information bits required to order N objects is $\log_2 N!$, where in this context $N = n_p + n_i$. Although much more is at stake in this situation than mere sorting, for errorless data TAILOR may be viewed as a sorting algorithm. It is gratifying that this theoretical minimum was closely approximated and often surpassed by all but one cell, namely, $n_p = n_i = 40$. For that condition, however, the obtained percentage of responses was .46, while the expected minimum was .25.

The measures of accuracy were the validities between true score and observed score. The validities for all conditions approached unity, deviating only to the extent that ties in true scores produced ties in the observed scores also.

Several technical enhancements have been suggested as a result of this study which have resulted in a substantial savings in CPU time. In addition, modifications to the decision-making process have enabled some reductions in the size of the program. Two versions of the program have been described—one in APL for individual testing (McCormick & Cliff, 1977) and a FORTRAN version which is adapted from the method used in the current study (Cudeck, Cliff, & Kehoe, 1977).

The next phase will be to carry out a second series of studies, again using the monte carlo method but with a more realistic model. Of the major alternatives available, the four-parameter model from Birnbaum (1968) seems especially promising for this purpose.

References

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. Lord & M. Novick, *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 procedure for interactive tailored testing. *Educational and Psychological Measurement*, 1977, *37*, 767–769.

Knuth, D. E. *The art of computer programming* (Vol. 1 & 2). Reading, MA: Addison-Wesley, 1973.

McCormick, D., & Cliff, N. TAILOR-APL: An interactive computer program for individual tailored testing. *Educational and Psychological Measurement*, 1977, *37*, 771–774.

McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 1947, *12*, 158–167.

Acknowledgment

Preparation of this article was supported by the Office of Naval Research, Contract N00014-75-C-0684, NR150-373.

Author's Address

Robert Cudek, Department of Psychology, University of Southern California, University Park, Los Angeles, CA 90007.