Evaluation of Implied Orders as a Basis for Tailored Testing with Simulation Data

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

Monte carlo research with TAILOR, a program using Implied Orders as a basis for tailored testing, is reported. Birnbaum's (1968) three-parameter logistic model was used to generate data matrices under a variety of simulated conditions. It was found that TAILOR typically required about half the available items to estimate for each simulated examinee the responses on the remainder. Validity of Tailored score with True score was found to be within a few points of True score with Complete test score. Increasing item discrimination affected the efficiency of the tailored test, but the procedure was little affected by any of a variety of other factors.

This research is based on an approach to tailored, i.e., computer-interactive, testing which differs from other testing methods both in its theoretical basis and its practical implementation. Most approaches to tailored or adaptive testing, e.g., Jensema (1974), Lord (1970; 1971), McBride (1977), Urry (1977), and Weiss (1976; 1977), are based on true score theory, employing items whose parameters have already been estimated on large pretest samples so that the individuals' true scores can be estimated by employing parametric, continuous true score models. The interactive strategies they employ are of several different types, but all follow this general mode. In contrast, the present method, called Implied Orders Tailored Testing, seeks only to define the order of examinees and the joint order of examinees and items according to a method whose basic rationale was given by Cliff (1975). A practical difference is that Implied Orders does not need to assume any pretest information on the items; item and person characteristics are estimated in parallel.

In previous reports on this research, a rationale for tailored testing (Cliff, 1975) and an example with errorless data (Cudeck, McCormick, & Cliff, 1979) have been presented. Implied Orders Tailored Testing, implemented through a program called TAILOR (Cudeck, Cliff, & Kehoe, 1977) has as its goal a simultaneous ordering of persons and items along the hypothetical ability continuum. It is based on the concepts of order theory, in which it is well known that the logical properties of an order are such that if certain of the relations among elements are known, then the remainder can be deduced from them by making use of the transitivity property, which characterizes orders. The general idea is that even an incomplete matrix of responses of persons to items can be used to deduce at

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 3, No. 4 Fall 1979 pp. 495–514

least some order relations between items which is relative difficulty. These order relations between items, in turn, can be used to predict the individual's responses to items not yet taken, therefore removing the necessity of administering those items. Cudeck et al. (1979) reported that TAILOR exactly recaptured the order between persons and items on the basis of approximately half the responses when the data used was errorless. The purpose of the present research was to evaluate the method in an artificial experiment in which Birnbaum's (1968) three-parameter logistic model was used to generate the data, thus determining its effectiveness under more realistically stochastic conditions.

TAILOR

Errorless Case

The mode of operation of TAILOR follows the general ideas suggested by Cliff (1975). The basic concept may be stated as follows: Suppose there are n persons to be tested and k dichotomous items that may be used in the tailored test. Then, at any stage in the testing process, there is an $n \times k$ matrix, **S**, of scores on the items, in which $s_{ij} = 1$ if the item was answered correctly and 0 if it was incorrect *or if it was not yet taken*. There is a second $n \times k$ matrix, $\tilde{\mathbf{S}}$, *in which* $\tilde{s}_{ij} = 1$ if the item was answered incorrectly, and 0 if it was correct or had not been taken. Then, there is a supermatrix containing both **S** and $\tilde{\mathbf{S}}'$, which is multiplied by itself:

$$\begin{bmatrix} 0 & \tilde{S}' \\ S & 0 \end{bmatrix}^2 = \begin{bmatrix} N & 0 \\ 0 & X \end{bmatrix}, \qquad [1]$$

where 0 is a null matrix, and an element n_{jk} of **N**, called the *integer item dominance matrix,* contains the number of persons failing item j and passing item k. Similarly, x_{ih} is the number of items passed by person i and failed by person h, and **X** is the *integer person dominance matrix.*

The elements of **N** may be used to infer the relative difficulty of the items. If the data are completely consistent, either n_{jk} or n_{kj} will be zero. If neither is zero, the magnitude of the difference between them implies the relative difficulty of the pair. Correspondingly, the elements of **X** may be used to infer the relative ability of pairs of persons.

This relative difficulty information may be used to estimate responses to items not yet taken. A matrix **N*** is constructed by setting $n^*_{jk} = 1$ if j is more difficult than k. Then $\mathbf{SN^*} = \mathbf{Y}$ is computed, and y_{ij} is the number of items *passed* by i which are *more difficult* than j. Similarly, $\tilde{\mathbf{Y}} = \tilde{\mathbf{S}}\mathbf{N^*}'$ is computed, and $\tilde{y}_{ij}$ is the number of items *failed* by i. With perfectly consistent data, one or the other of y_{ij} and $\tilde{y}_{ij}$ will be zero. If y_{ij} is not zero and $\tilde{y}_{ij}$ is, this implies that i would answer j correctly if he/she took that item. If the reverse is true, it implies he/she would get it wrong. In either of these cases, the appropriate element of **S** or $\tilde{\mathbf{S}}$ is set equal to 1, even if the person has not taken the item. In some cases both y_{ij} and $\tilde{y}_{ij}$ may be zero, in which case more items must be taken if this response is to be implied. In this way an incomplete (tailored) items response matrix may be completed if the data are perfectly consistent.

This is the general idea of the mode of operation of TAILOR. At any given time, some items have been taken by some persons. The foregoing process, modified to take account of the fallible nature of empirical data, is used to infer additional responses. Then, additional items are administered, but the new items are those whose responses cannot be inferred from those already taken. This process continues until all responses are made or implied.

Two additional aspects of TAILOR may be noted. First, $(\mathbf{N}^*)^2$ may be computed and used to infer additional, indirect item dominance relations; these are a kind of second-order item dominance relation. In fact, this process could be continued for further powers, but TAILOR uses only $\mathbf{N}$ and $(\mathbf{N}^*)^2$ as currently designed. Second, the whole process could be carried out with the roles of $\mathbf{N}$ and $\mathbf{X}$ reversed; TAILOR includes this option for use in case the number of persons is less than the number of items. To reduce awkwardness of presentation, the description in this paper will emphase item dominances, although this is not required.

Two additional major aspects of the program were crucial to its successful operation. One was the introduction of a "significance test" to apply with fallible data in the comparison of n_{jk} to n_{kj} and y_{ij} to $\tilde{y}_{ij}$. The other was a heuristic rule for deciding which item was most appropriate for a given person at a given time.

"Significance Test"

The fact that fallible data will often show both n_{jk} and $n_{kj} \neq 0$ can be dealt with by setting $n^*_{jk} = 1$ if $n_{jk} > n_{kj}$. However, when $n_{jk} = n_{kj}$, the true order between j and k will be in doubt. In TAILOR, the doubt is resolved by one of two "loose significance tests." The major one corresponds to comparing frequencies n_{jk} to n_{kj} or y_{ij} to $\tilde{y}_{ij}$ by a binominal probability (McNemar's test; McNemar, 1969) and rejecting the null with a one-tailed alpha level of .22. For example, values of n_{jk} and n_{kj} of 2 to 0 and 3 to 1, respectively, lead to rejection by this rule and therefore imply that j is more difficult than k.

The second decision rule concerns instances in which the frequencies are 1 and 0. If the information is sparse (i.e., early in the testing), most of the frequencies compared are either 0,0 or 1,0; but it is still necessary to obtain some information from the latter in order to differentiate between items. However, when the information is less sparse, frequencies of at least 1,0 are almost bound to occur. Thus, the decision whether a 1,0 frequency is likely to reflect a true ordinal relation depends on how sparse the information is. The nature of this decision rule is described in the Appendix.

Net-Wins Scores

Since the basic premise of the method lies in trying to form a joint order of persons and items, it is natural that information about this joint order be used to decide which item a person should take at any given time. This was done by means of a *Net-wins score* for persons and items. At any given time each person is assigned the item which is currently nearest him/her in the joint order furnished by the Net-wins score, subject to some restrictions concerning the number of persons who can take an item at the same time.

The Net-wins score requires the estimation of the person dominances as well as item dominances and actual and implied responses. The person dominances are derived from the matrices of actual and implied responses; let $\mathbf{S}^*$ and $\tilde{\mathbf{S}}^*$ be matrices of the actual plus implied correct and incorrect answers, respectively. Then $\mathbf{S}^*\tilde{\mathbf{S}}^{*\prime}$ is an $n \times n$ matrix whose entries are the number of actual or implied items that the row person gets correct and the column person incorrect. The symmetrically placed entries in this matrix can be compared by the "significance-testing" procedure described earlier; and, in effect, a binary person-dominance matrix $\mathbf{X}^*$ is constructed from $\mathbf{S}^*\tilde{\mathbf{S}}^{*\prime}$. The Net-wins score of a person is then the number of items he/she dominates (actual or implied rights) plus the number of persons he/she dominates (from $\mathbf{X}^*$) minus the number of each which dominate him/her. The first two components are given by the number of elements in his/her row of $\mathbf{S}^*$ and the number in his/her row for $\mathbf{X}^*$, respectively; and the latter two are given by the number of elements in his/her *column* of

$\tilde{\mathbf{S}}^{*\prime}$ and $\mathbf{X}^*$, respectively. An item's Net-wins score has exactly the same definition. It is the number of items and persons it dominates minus the number which dominate it. The former are the number of elements in its row of $\mathbf{N}^*$ plus the number in its row of $\tilde{\mathbf{S}}^{*\prime}$; the latter is the sum of the elements in its column of $\mathbf{N}^*$ plus the number in its column of $\mathbf{S}^*$. Thus, the Net-wins for both items and persons are derived in exactly the same way from binary information concerning ordinal relations.

Testing Rounds

TAILOR carries out the above sequence of operations repeatedly, operating on a group of items simultaneously. It starts by giving every person an item at random. It then calculates the small number of relative difficulty relations possible with such little data and enters them in a temporary $\mathbf{N}^*$ matrix, which it then uses to calculate $\mathbf{S}^*$ and $\tilde{\mathbf{S}}^*$, which in turn are used to calculate $\mathbf{X}^*$. At this stage there are many ties in the order; but the fact that TAILOR is already coming to some tentative relative difficulty decisions means that there are more than three possible scores after only two responses per person, and it is possible that there already are some implied responses.

Each person is then assigned one of the items closest to him/her in the joint order. He/she is, of course, not assigned any item to which his/her response is already implied, and a restriction is placed on the number of persons who can be assigned the same item.

The new round of responses is entered into the $\mathbf{S}$ and $\tilde{\mathbf{S}}'$ matrices. The latter are used to calculate a new integer item dominance matrix $\mathbf{N}$, and from this in turn flows new versions of $\mathbf{N}^*$, $\mathbf{S}^*$, $\tilde{\mathbf{S}}^*$, and $\mathbf{X}^*$. These provide the basis for a new set of item assignments, and the whole process starts again. The mode of operation means that ordinal relations which exist in any of the binary matrices at one stage of the process may be cancelled and even reversed at later ones. Thus, no permanent harm is done by the procedure when it comes to erroneous conclusions on very little data in the early stages. Even then, most of the decisions it makes are correct, so they aid in correct assignment of items to persons.

The process terminates when there is an actual or implied response by every person to every item; the $\mathbf{S}^*$ and $\tilde{\mathbf{S}}^*$ matrices are complete. This takes place on the basis of fewer responses for some persons than for others, but the variation is not very great.

Other versions of the same general procedure exist. For example, TAILOR-APL (McCormick & Cliff, 1977) is designed to test single subjects, with ordinal information therefore accumulating as more and more individuals are tested, so that later examinees take fewer items than earlier ones. For practical reasons, the "lock-step rounds" approach was also revised in later versions of TAILOR itself. Nonetheless, it was this version which was used in the study described here. Results would be expected to be closely applicable to other versions of the same basic procedure.

The model underlying this tailored procedure is then a rather simple one. It states:

1. If a person frequently fails some items that are easier than item j and seldom answers correctly items that are more difficult than j, then he/she will probably fail j; if the reverse holds, he/she will pass it.
2. An item j is easier than an item k if the proportion of persons passing j and failing k is greater than the proportion doing the reverse.

The McNemar's test and the probability test which are used here are adjuncts to this model. They constitute a decision rule for deciding whether Statement 1 holds for the pair of items and whether Statement 2 holds for a pair of items over the population of persons.

Effectiveness of TAILOR

The present study was designed to investigate whether the procedure described above could be expected to be effective with real data and to study some factors which were likely to affect its efficiency. For these purposes, a monte carlo simulation has many advantages over real-data studies, although the latter are necessary for the ultimate confirmation of findings.

Several variables seemed relevant to TAILOR's operation, but the one that was the major source of concern was the degree of consistency of the responses. Completely consistent data were bound to give good results, as had already been shown (Cudeck et al., 1979), and no benefit is to be expected from making inferences from responses which are completely random; but it is difficult to predict what would happen at intermediate degrees of consistency. Similarly, it would be important to know the effect of the presence of guessing, since this would be likely to introduce additional random components. Situational variables such as the number of items in the test and the number of examinees being tested might also be important. The difficulty distribution of the items in terms of whether they were centered in the middle of the ability distribution and whether they were relatively constant or fairly variable in difficulty might also have an effect on how TAILOR operated. This study investigated the effect of these variables by varying parameters of the simulation. The Birnbaum (1968) model was used as a device for the convenient generation of data, even though the present model represents quite a different philosophy, since it allows the ready manipulation of various factors thought likely to affect the operation of TAILOR.

As outlined above, TAILOR produces a final rank order of persons from the partial response matrix using the Net-wins score at the point where the procedure terminates. The major question concerns the accuracy of this order as a predictor of the "true" person order and how closely this accuracy approaches that which would have been derived from the complete response matrix, in which every person takes every item. Thus, primary interest was in the overall level of this accuracy for some sets of conditions that might be typical of the range of conditions of actual testing. Of secondary, but not negligible, interest was the overall level of efficiency in computer time and the factors which affect it; for a procedure which is psychometrically effective but required enormous computing investment to achieve that efficiency would not be of great practical applicability. Therefore, this investigation was principally concerned with the correlation between obtained scores and true scores, and also with the amount of computer time used.

Method

Birnbaum Model

The Birnbaum (1968) model postulates a probability, p_{ij}, of person i correctly answering item j as

$$p_{ij} = c_j = (1 - c_j) \frac{1}{1 + \exp\left[-1.7a_j\left(\Theta_i - b_j\right)\right]} \qquad [2]$$

where

a_j = item discrimination, $j = 1, 2, \ldots k$,
b_j = item difficulty
c_j = probability of chance success, and
Θ_i = person ability, $i = 1, 2, \ldots n$.

The parameter vectors used in this data were assumed to have the following distributions:

$$\underline{a} = N(\mu_a, \sigma_a) \quad [3]$$

$$\underline{b} = N(\mu_b, \sigma_b) \quad [4]$$

$$\underline{c} = \text{constant} \quad [5]$$

$$\underline{\Theta} = N(0, 1) \quad [6]$$

The normal variates were generated from a sequence of uniform random numbers in

$$V_g = \sigma_g \left[\sum_{k=1}^{12} U(k) - 6\right] + \mu_g \quad [7]$$

with

V_g = the normally distributed vector of item or person characteristics,
σ_g = the parameter standard deviation,
μ_g = the parameter mean, and
$U(k)$ = a function for uniform random numbers in the 0, 1 interval, based on a method from Knuth (1973).

A run in the simulation began by specifying the number of persons and items to be tested and selecting values for the parameters of the distributions of *a*, *b*, and *c*. Then the required numbers of values for the item parameters and for Θ were sampled according to Equation 7. These were inserted in Equation 2, and an additional random number $U(k)$ was generated. An element s_{ij} of the score matrix **S** was set equal to 1 if $p_{ij} > U(k)$, and zero otherwise. A complete score matrix was prepared in this fashion. Then "presenting item *j* to person *i*" consisted of looking in this score matrix to see if the response was correct or incorrect.

A given condition was replicated by repeatedly sampling the item parameters from the population specified by the condition, generating a new sample of ability scores (Θ) and a new score matrix by inserting them in Equation 2 and then generating a new $U(k)$. Therefore, the actual characteristics of the items varied from replication to replication according to basic sampling principles. This was felt to provide a more realistic simulation than fixing the item parameters at specified values or making them constant from replication to replication.

Design

The item parameters then had distributions, and the mean and variance of the distributions could be considered determiners of the characteristics of the test. Thus, provided $c = 0$, μ_a and σ_a de-

fined the internal consistency of the test, the former the overall level of consistency and the latter the degree to which items were equally consistent or varied. Setting c greater than zero would simulate a multiple-choice test, and the value selected could be used to simulate different numbers of distractors and/or whether they tended to be attractive. The value of σ_b determined whether the range of item difficulties was small or large, and setting $\mu_b \neq 0$ would allow the difficulty distribution to be made nonoptimum.

The fact that the item parameters were sampled from their specified populations separately for each replication meant that there would be some variation in them. However, in the experiment, the size of the manipulations of the parameters were selected to be large relative to these residual sampling effects. Nevertheless, some sampling deviations from the population values set in the conditions remained.

The variables whose influences were investigated and the labels used for them were the following:

1. *Number of examinees assumed to be tested at the session* (*Persons*). Included to test the procedure's sensitivity to the size of the examinee sample.
2. *Number of items in the item pool* (*Items*). Included to examine the relation between TAILOR's effectiveness and the size of the item pool.
3. *Mean item discrimination index* (μ_a). Included to investigate the effect of the basic discriminatory power of the items.
4. *Variation in discrimination* (σ_a). Investigated the effect of heterogeneity in discriminatory power.
5. *Mean difficulty* (μ_b). Setting $\mu_b \neq 0$ would show whether the process was affected by the appropriateness of the average difficulty level of the items.
6. *Variation in difficulty* (σ_b). Manipulation of σ_b showed whether TAILOR worked better with small or large variation in difficulty.
7. *Chance probability* (c). This showed the extent to which TAILOR's effectiveness would be affected by guessing.

With this number of variables, it was not possible to vary all of them simultaneously or over a wide range of values. Instead, two or three values of each were selected on the basis of realism and practicality, and small factorial designs involving two or three of the variables were constructed. In this way, the main effects and certain first- and second-order interactions could be studied, but higher order interactions could not. The combinations selected were chosen on the basis of their expected importance, their practicality in a study of this kind, and their resemblance to empirical conditions.

A total of 28 different conditions were run, as shown in Table 1. The assumed number of subjects was 10, 25, or 40, with the great majority of the data coming from the latter two values. The number of items was assumed to be either 15 or 25. These rather small numbers of persons and items were chosen for reasons of economy. Mean discrimination was assumed to be either $\mu_a = 1.0$ or 2.0, except in two cases where it was .5. The standard deviation of discrimination was usually assumed to be 0, but also took on values of .2 or .4. The mean population difficulty was usually assumed to be $\mu_b = 0$, equal to the mean ability, but was also set at +.5 (average difficulty one half sigma above the score mean) for two cells. Similarly, the standard deviation of difficulty was usually 1.0, perhaps a rather large value, and was 2.0 for two cells. The chance parameter was usually $c = 0$, but also took on values of .1 or .2.

The basic part of the design was a $2 \times 2 \times 2$ factorial design with 25 or 40 persons, 15 or 25 items, and discrimination of 1.0 or 2.0. The other parameter variations were usually made singly in combi-

Table 1
Characteristics of Samples of Score Matrices Generated by Latent Trait Models

Condition Number	Sample Characteristics						
	Persons	Items	μ_a	σ_a	μ_b	σ_b	c
1	10	25	1.0	0	0	1.0	0
2	10	25	2.0	0	0	1.0	0
3	25	15	.5	0	0	1.0	0
4	25	15	.5	0	0	1.0	0
5	25	15	1.0	0	0	1.0	0
6	25	15	2.0	0	0	1.0	0
7	25	15	1.0	0	.5	1.0	0
8	25	15	1.0	0	.5	1.0	0
9	25	15	1.0	0	0	2.0	0
10	25	15	2.0	0	0	2.0	0
11	25	15	1.0	.2	0	1.0	0
12	25	15	2.0	.2	0	1.0	0
13	25	25	1.0	0	0	1.0	0
14	25	25	2.0	0	0	1.0	0
15	25	25	1.0	0	0	1.0	.1
16	25	25	2.0	0	0	1.0	.1
17	25	25	1.0	0	0	1.0	.2
18	25	25	2.0	0	0	1.0	.2
19	25	25	1.0	.2	0	1.0	0
20	25	25	2.0	.2	0	1.0	0
21	40	15	1.0	0	0	1.0	0
22	40	15	2.0	0	0	1.0	0
23	40	15	1.0	0	0	1.0	.2
24	40	15	2.0	0	0	1.0	.2
25	40	25	1.0	0	0	1.0	0
26	40	25	2.0	0	0	1.0	0
27	25	25	2.0	.4	0	1.0	0
28	25	15	2.0	.4	0	1.0	0

nation with different levels of one or two of these main parameters. Each row of Table 1 defines a set of conditions for a sample of score matrices generated by the Birnbaum model.

Procedure

Five sample score matrices were generated according to the parameters of each line of Table 1. In addition to the true score Θ that was used to generate the data, there are two scores for each person in a given sample. One was the Net-wins score from the tailored simulation, and the other was number-

correct score on the complete test. These will be referred to as Tailored scores and Complete scores, respectively. Note that they are not experimentally independent, because the responses on which the Tailored score was determined was a subset of the responses determining Complete score. A number of statistics were calculated from these, and some of them were important to the later analysis.

The correlations of True score with Tailored score and Complete score were computed. The two correlations with True score were the validities of Tailored and Complete scores. These *correlations* were used as a major dependent variable and covariate, respectively. The reasoning was that agreement with True scores is the major quality desired from a tailored testing procedure, and the expectation was that variations in Complete score validity would be the major source of influences to be controlled. Both Pearson and rank-order (tau_b) coefficients were computed. The Fisher Z-transformation was not used because previous experience indicated that it has little effect on results unless the mean and the variance of the correlations are both large, and rarely even then.

Other dependent variables that were used reflected efficiency and cost of the procedure. In particular, the ratio of actual responses to total possible responses in a given score matrix was clearly relevant, as was the amount of computer-processing time used per run. The effects of the independent variables on these were assessed and the general overall levels were determined.

The analysis of the effects of the independent variables was carried out using regression analyses or analyses of variance (ANOVA) and analyses of covariance (ANCOVA). The correlation of True score with Complete score for the same data was used as the single covariate. This permitted the assessment of the degree to which there were effects over and above that of the basic consistency of the data. In addition to the numerous small analyses, a regression analysis with all the data combined into a single score matrix (very much like Table 3) and with the main effects as independent variables was carried out, both with the covariate and without.

Results

Principal Findings

The major results are included in Table 2, which shows the mean values of the important variables under conditions defined in Table 1. Table 3 gives the correlations among the main variables in the study. Here, both Pearson and tau correlations of True score with Tailored score are given, as well as True score with Complete score. The percentage of responses under TAILOR and the amount of CPU time used are likewise presented.

The bottom line of Table 2 shows the mean of each variable. These means provide a quick summary of the major findings. On the average, TAILOR presented 55% of the items to each person. The mean validity correlations for the Tailored scores averaged .757 (tau) and .889 (r). When compared to the Complete score validities of .810 and .926, the results indicate that TAILOR performed quite well over the range of conditions. The means represent averages across conditions that affect these variables, and some of the effects were substantial. Therefore, their precise values should be viewed rather tentatively. This optimistic picture can be more closely examined by discussing the major sources of variance, reflected by the appreciable deviation from these means.

Influences on Proportion of Responses

The major influences on the proportion of responses appeared to be the number of items and persons, where there was an inverse relation. Table 4 shows the mean proportion of responses for the 2 × 2 × 2 subset of the study mentioned earlier, plus the results for the 10-person, 25-item condition.

Table 2
Cell Means of Major Dependent Variables

Condition Number	Complete Data		Tailored Data		Percent of Responses	CPU (Per Person)
	Tau (b)	Pearson	Tau (b)	Pearson		
1	.769	.941	.724	.850	.590	3.64
2	.880	.955	.846	.946	.577	3.54
3	.745	.876	.636	.810	.615	.54
4	.621	.803	.480	.686	.556	.58
5	.803	.920	.750	.896	.609	1.34
6	.860	.954	.818	.934	.580	1.14
7	.771	.916	.734	.898	.599	1.38
8	.863	.932	.822	.922	.577	1.22
9	.662	.869	.600	.810	.552	1.11
10	.865	.958	.854	.952	.571	1.16
11	.771	.924	.728	.880	.603	1.32
12	.863	.938	.836	.928	.592	1.36
13	.809	.924	.738	.878	.514	3.62
14	.868	.965	.884	.946	.516	3.59
15	.826	.947	.754	.904	.538	3.39
16	.836	.946	.824	.920	.506	3.82
17	.788	.918	.696	.866	.555	4.70
18	.821	.913	.760	.896	.502	3.84
19	.821	.945	.758	.884	.506	3.38
20	.890	.968	.864	.954	.490	3.53
21	.766	.904	.710	.868	.543	1.44
22	.855	.932	.836	.924	.548	1.41
23	.682	.843	.560	.734	.567	1.57
24	.773	.907	.706	.870	.571	1.54
25	.842	.956	.772	.920	.471	3.14
26	.892	.956	.852	.940	.441	3.14
27	.868	.962	.846	.948	.492	3.60
28	.836	.940	.814	.924	.585	1.23
Mean	.810	.926	.757	.889	.549	2.39

Here, the strong effect of items and persons was quite apparent, but the effect of discrimination was weak or nonexistent and was nonsignificant in the ANOVA for these data. For the 25-item, 40-person cell, the mean proportion of responses was .456, indicating that in large-scale testing, fewer than half the items will be needed.

Table 5 shows the results of a regression analysis for all 28 conditions (n = 140) of proportion of responses on all three variables, treated as main effects. There, the F-column shows the significance of the regression weight for the individual variables; items and persons were clearly significant, but discrimination was not. Thus, the proportion of items used decreased sharply with the number of items in the test and also significantly with the number of persons, but was not affected by anything else.

Table 3
Correlation Matrix of Principal Independent and Dependent Variables from Monte Carlo Studies

		1	2	3	4	5	6	7	8	9
1.	Number of Persons									
2.	Number of Items	-.26								
3.	Mean Discrimination	.03	.16							
4.	Mean Chance	.02	.14	.06						
5.	Mean Difficulty	.08	-.26	.03	-.14					
6.	Tau: Abiliity/Complete Score	-.06	.33	.60	-.18	.02				
7.	r: Ability/Complete Score	-.13	.40	.53	-.21	-.01				
8.	Tau: Ability/Tailored Score	-.10	.31	.70	-.25	.86	.80			
9.	r: Ability/Tailored Score	-.09	.29	.65	-.22	.08	.82	.83	.92	
10.	Proportion of Responses	-.28	-.57	-.22	-.05	.20	-.22	-.21	-.19	1.0

Table 4
Influence of Number of Persons and Items, and Average Discrimination on Proportion of Items Used

10 Persons: 25 Items		25 Persons: Items			40 Persons: Items		
Discrimination		15	25	Mean	15	25	Mean
1.0	.590	.609	.514	.562	.543	.471	.507
2.0	.577	.580	.516	.548	.548	.441	.494
Mean	.584	.594	.515	.555	.546	.456	.501
	n=2		n=20			n=6	

Validity of Tailored Scores

The success of a tailored testing scheme is primarily measured by the ability of the Tailored scores to substitute for the Complete test scores. In a monte carlo study such as this, in which a True score is available, correlation with True score would appear to be the most appropriate criterion by which to judge the success of TAILOR.

As noted earlier, the Tailored validities were only slightly less than the Complete score validities but were quite variable, ostensibly as a function of various experimental parameters (see Table 2). A number of analyses were made to attempt to identify the characteristics that affected Tailored validity.

The primary effect was due to the consistency of the original sample of data. This is illustrated graphically in Figures 1 and 2, which plot mean Tailored correlations as a function of mean Complete correlation for each of the 28 conditions. Figure 1 shows tau; and Figure 2, *r*.

Treating each of the five replications under each condition separately, the correlation between Complete and Tailored validities were .86 and .83 for tau and *r*, respectively, showing a strong dependence of Tailored validity upon the consistency of the data. The slopes of the regression lines in the figures are greater than unity, showing that a variable having an effect on the validity of a complete test will have an even greater effect on the validity of the items in a tailored format.

Influences of Independent Variables on Validity

A variety of regression analyses and analyses of variance and covariance were performed to identify influences on validity. Tailored tau, the rank-order correlation of Tailored score with True score, was the dependent variable in most of these analyses. Some of the analyses were duplicated with both dependent variables, *r* and tau; and the results were never in conflict. Tau was emphasized for two

Table 5
Regression of Proportion of Items on Persons, Items and Discrimination

Variables	b	Beta	F	R^2
Items	-.0076	-.690	128.05	.327
Persons	0.0036	-.459	56.69	.524
Mean Discrimination	-.0101	-.092	2.38	.532

Figure 1

Relationship between Complete Test Tau and Tailored Test Tau for 28 Conditions

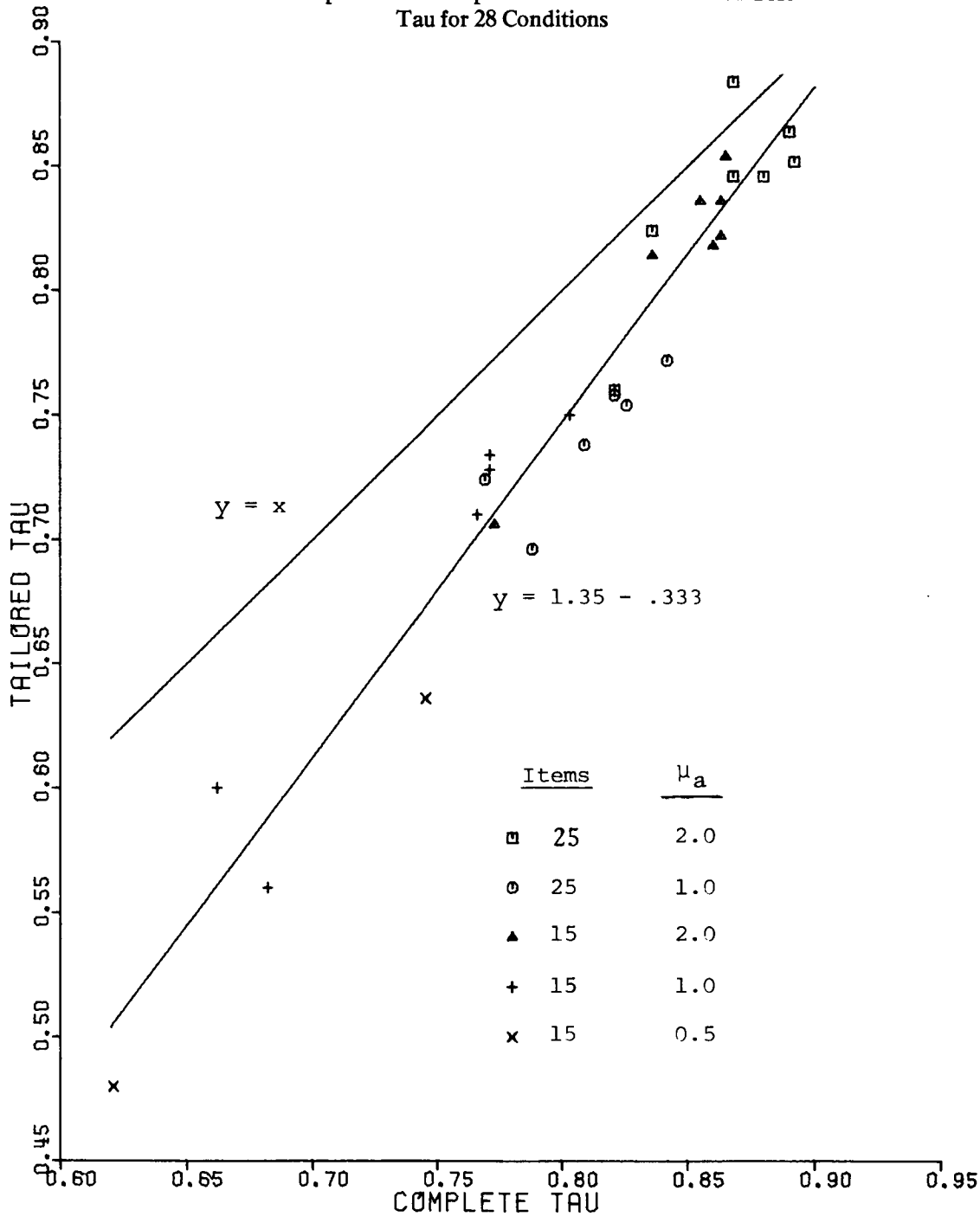
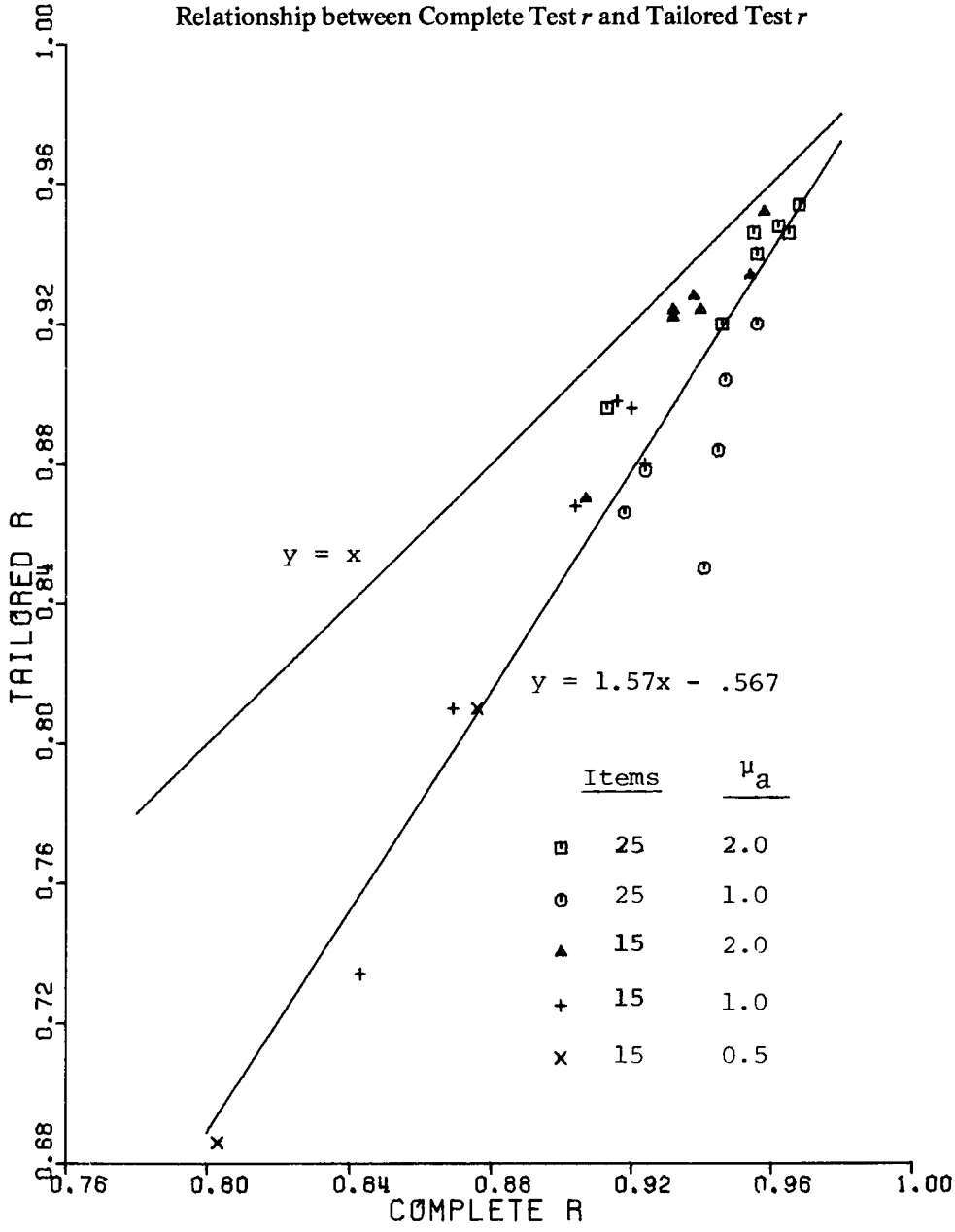


Figure 2

Relationship between Complete Test *r* and Tailored Test *r*



reasons. One is the ordinal emphasis of TAILOR, making the rank-order agreement more appropriate. The other is of a more pragmatic nature. There is more variability in tau, and it is farther from the limit of 1.0, whereas the range of *r* is toward high values where variability is limited. Furthermore, tau showed slightly greater sensitivity to the independent variables, viz., the .86 versus .83 validity correlations for Complete test score. Thus, the effects showed slightly more clearly with tau.

Regression analyses based on the values of the parameters in Table 1 for all 28 conditions were performed, treating Tailored tau as the dependent variable. Analyses of variance of subsets of that data were also performed, as were analyses of covariance treating Complete tau as the covariate. These showed only one significant interaction; therefore, the regression analysis, which includes the main effects of all the independent variables, provides a valid summary. The analyses of variance and covariance will be presented also, however, in a few instances.

Since the validity of the Tailored score is dependent on the validity of the original data, the validity of the Complete data score serves as a limiting value on Tailored validity. The validity of Complete score from the following analyses was itself found to be affected by three manipulated variables and by a random effect that is dependent on the actual sampling of items (see Table 3). The three manipulated variables were the mean discrimination index for the population of items, the probability of chance success, and the number of items.

Table 6 shows the results of a stepwise regression analysis of Tailored tau on predictors. The upper section shows the results when Complete tau was included as a predictor; and the lower, when it was not. The "significant variables" are those which significantly increased the multiple R as they were included. After they were included, none of those remaining were found to add significantly to the multiple R by any plausible interpretation of "significant."

Complete tau was the first variable entered in the regression equation, accounting for 74% of the variance of all 140 observations (5 replications in 28 conditions). However, two of its causes—mean discrimination and chance probability—still added significantly to the multiple R^2, contributing 5% and 2% of the variance, respectively. All three had weights of the expected sign: High Complete tau and higher mean discrimination led to higher Tailored tau, while greater chance probability led to lower.

When only the manipulated variables were included, the third influence on Complete test validity also became significant. Now, mean discrimination accounted for almost half the variance (19%), and chance probability contributed another 8%. Items added a final 5%. The total percentage of variance accounted for was 63%, rather than 82% as it was when the consistency of the actual data set was included. The number of persons, the variability in discrimination of the items, and the mean and standard deviation in difficulty were not significant influences on validity with the levels of parameters used here. The level of precision here (witness the small proportion of variance for items, which was nevertheless significant) was such that these variables must only have small effects, if any. The lack of negative effects for numbers of persons and items is particularly interesting in view of the fact that a smaller proportion of responses were used as these increase.

Table 6
Significant Predictors of Tailored Tau

Variables	b	Beta	F	R^2
Including Complete Tau				
Complete Tau	.8124	.6400	183.64	.743
Mean Discrimination	.0585	.3270	49.32	.796
Chance Probability	-.2366	-.1525	16.15	.818
Manipulated Variables Only				
Mean Discrimination	.1215	.6794	162.94	.490
Chance Probability	-.4910	-.3166	35.68	.572
Items	.0053	.2362	19.39	.625

Subsidiary Findings

The 28 conditions contained a number of subdivisions that represent orthogonal designs with two or three of the manipulated variables. These allowed the investigation of a number of first-order, and a few second-order, interactions among the variables, as well as the highlighting of the main findings. That is, there were a number of small factorial designs contained in the total set; the other variables were held constant at some particular combination of levels.

Analyses of variance (and analyses of covariance with Complete tau as the single covariate) of these subdesigns gave essentially the same picture as the regression analyses just described, but in some cases they pointed up unique findings or gave a more direct impression of the magnitude of effects. The particulars will not be presented here; interested readers are referred to Cliff, Cudeck, and McCormick (1977) for an examination in detail. Some specific findings will be briefly noted, however. The following summary remarks can be verified by examining the data provided in Table 2.

When mean discrimination was changed from 1.0 to 2.0, Tailored tau was raised about .10, other factors held constant. Lowering it to .5 had an approximately equivalent effect. This is largely because the reliability of the basic data is closely related to discrimination. Changing from 0 to a .2 guessing probability had almost the same effect as going from 1.0 to 2.0 discrimination. On the other hand, the degree of variability in discrimination of the items in a pool did not show any effect, at least over the moderate ranges studied here. These findings echo those from the regression analyses; the only significant interaction will be mentioned later.

Somewhat surprisingly, no effect for the number of persons was found, even though the sample size was reduced to 10 in two conditions. As shown in Table 4, TAILOR presented a somewhat larger proportion of the items with fewer subjects, and apparently this compensated for the greater uncertainty of the process with very small samples.

The findings with respect to the number of items were similarly negative. With a larger item pool, TAILOR tended to give slightly more valid scores, but the differences were not quite significant. Even these differences were largely accounted for by the fact that TAILOR gave a slightly larger number, although a small fraction, of items when the item pool was larger. The quality of its performance, thus, was not sensitive to the size of the item pool.

Most of the 28 conditions assumed that the item difficulty distribution was the same as that of ability, but there were some exceptions. When the difficulty distribution was centered a half-sigma away from the ability distribution, there was no effect. Thus, TAILOR was not sensitive to the exactness of the difficulty match. There was no main effect for the variance of difficulty either, but this variable did furnish the only significant interaction which was discovered. It turned out that highly variable difficulties gave higher validity when the discrimination index was high, but moderately variable difficulties were better when discrimination was merely good. However, this effect also occurred with the Complete validities and vanished in the ANCOVA. That is, the effect appeared to be a property of test items in general, rather than a methodological result of using TAILOR. Unfortunately, a $\sigma_b = 0$ condition was not run to enable testing of the limits on this finding.

Overall, the results of these subsidiary analyses supported the picture shown by the regression analysis. There was a substantial robustness of TAILOR with respect to the variables manipulated. The variables that affected the validity of TAILOR scores were those that affect the validity of conventional test scores.

Computer Time

A real-time system is only useful if it can operate efficiently, and if a computerized testing system is to be adopted, it cannot be too costly. The amount of central processing unit time (CPU) used in

each computer run was recorded as part of the operation of the program. A few conditions that were anomalous for technical programming reasons were deleted, and the average CPU for each of the main combinations of persons and items were computed. These are given in Table 7, where it is apparent that both had a substantial effect, particularly items. Prorated across persons, the data indicate that about 4 seconds of CPU was expended per subject with a pool of 25 items. This is, admittedly, on a highly efficient IBM 370/158 installation, but at charges which were about five cents per CPU second, computing costs do not seem to be a major factor. It should be noted as well that a good part of the computer time was used for overhead routines used to monitor the process, which would not be included in an operational version of the program. Furthermore, substantial increases in program efficiency were instituted since these data were gathered, and computing costs seem to be continuing their historic decline, rather than leveling off. Therefore, it is not foreseen that computing will be a major expense, even with item pools of substantially larger size.

Discussion

How much does a tailored test save? One answer to this can be found by comparing the reliability of a tailored test to one which is simply shortened to an equivalent length. Alternatively, given the reliability of tailored and complete tests, the Spearman-Brown formula (Lord & Novick, 1968, p. 112) can be solved for the length factor and can be compared to the actual proportion of items asked in the tailored version.

The latter was done, starting by squaring the validities to obtain a reliability estimate. The Pearson correlations were used for this purpose. The most representative case is the data for 25 items and 40 persons, with discriminations of 1.0 and 2.0—Conditions 25 and 26 of Table 2, where Tailored validities were .920 and .940, respectively, and the Complete validities were .956 and .965. The relation of .965 to .940 corresponds to using a test 78.4% as long, whereas in actuality only 44.1% of the responses were used by TAILOR. That is, a test that required only an average of about 11 responses acted like a complete test with nearly 20 items. The corresponding data with discriminations of 1.0 were not quite as favorable, but still were encouraging. Here, the validities of .956 and .920 corresponded to using a test 72.6% as long, whereas in fact TAILOR used 47.1% of possible responses. Here, 12 responses acted like an 18-item test. More exactly, the ratio of actual responses to lengths estimated from reliability was 1.778 for the 2.0 discrimination item pool, and 1.541 for 1.0 discrimination.

The corresponding calculations for the smaller item pools were also favorable, but not as much so. For 15 items the ratios of the number of responses to lengths estimated from reliabilities were 1.612 for 2.0 discrimination and 1.257 for 1.0, primarily because of the larger proportion of items used. The relations between the results for 25 and 15 items suggests that the savings for item pools of a more realistic size will be even more substantial. That is, the process becomes relatively more effi-

Table 7
Central Processing Time in Seconds by Items and Persons

Items	Persons	
	25	40
15	31.3	59.5
25	94.7	143.9

cient as the number of items increases because the items that are used are, on the average, closer to the person's ability level.

The question which may be raised with respect to the usefulness of TAILOR has to do with its sensitivity to the consistency of the data. A discrimination parameter of 1.0 is near the upper reaches of what can be expected with real items, corresponding to item-ability biserial of .707 with free-answer items (Urry, 1974). However, even the results for Condition 17, which simulated multiple-choice items with 5 alternatives, were still fairly good. This, coupled with the fact that validity held while the proportion of responses used decreased with the size of the pool and thus resulted in greater efficiency, suggests that TAILOR would operate successfully with only moderately discriminating items, provided that the pool was large. Thus, TAILOR might provide a reasonable method in a variety of situations.

Some aspects of the design of the present study may deserve comment. One is that the main measures of the procedure's effectiveness were actual correlations between True scores and scores on a subset of responses. Thus, the correlations were not confounded with assumptions concerning the accuracy of the model, as would be the case if standard errors of estimates of ability had been used. Also, the purely ordinal model here was shown to work well, even though the data were generated by means of a parametric true-score model. The study is therefore a relatively unconfounded test of the procedure.

One serendipitous finding may be noted: For both Complete and Tailored scores, a highly variable distribution of difficulty gave more valid scores than a moderately variable one when the discriminations were high. The reverse was true if they were less high. This is presumably a manifestation of a phenomenon related to the attenuation paradox (Loevinger, 1954), suggesting a trade-off between discrimination and difficulty variability, which may bear exploration in more detail. Concepts of interitem redundancy and uniqueness as determiners of the usefulness of items, as proposed by Cliff (in press), may well be relevant here.

Summary

It appears that the TAILOR procedure worked quite well under a variety of circumstances. Without any pretesting it arrived at a reasonable approximation to the Total score on a test, using about half the items. Moreover, the percentage decreased with the number of persons and items without a concomitant loss in validity. The major determinant of the validity of the Tailored score was the validity of the item responses on which it is based. The Tailored score was somewhat more sensitive to influences of consistency, such as mean discrimination and chance than the Total score. It was relatively robust with respect to variations in a variety of parameters, and TAILOR is computationally efficient enough for practical use, provided that the items are of the levels of quality used here.

Appendix

This test assesses the probability that elements distributed at random in two vectors will correspond on the basis of chance alone. Suppose a vector has n elements, n_j of which are 1 and the remainder 0. Suppose a second n vector has n_k 1's and the rest 0. If the n_j 1's are scattered at random in the first vector and the n_k 1's are scattered at random in the other, when the two vectors are laid side by side, what is the probability that none of the 1's in the one vector are matched by 1's at corresponding places in the other? If the complement of this probability is found, this is the probability of at least one pair of 1's with the same index in the two vectors, i.e., a frequency that is not 0,0. Ob-

viously, if $n_j + n_k$ is greater than n, there must be at least one match. If not, the probability of 0 matches is given by the following formula, where n_j is greater than n_k:

$$p(0) = \frac{\binom{n - n_j}{n_k}}{\binom{n}{n_k}} \qquad [A.1]$$

If $p(0) \geq .5$, this implies that a random match is unlikely to occur; therefore, the observed 1,0 frequency probably represents real information about the item order. If $p(0) < .5$, the probability is considered too great that the matching elements occurred by chance and that insufficient order information exists.

These standards will clearly seem incautious to anyone raised in the .05 to .01 tradition of significance testing. Two things should be borne in mind. Most of these implications of order are subject to reversal on the basis of later evidence, so the decisions are not irrevocable. Second, there is not the same payoff matrix here as that underlying traditional hypothesis testing. Particularly at the early stages, the penalty for concluding that there is not a difference in difficulty when one actually exists is as large as the penalty for concluding that there is a difference when in fact there is not. In fact, a good deal of exploratory simulation work forced the abandonment of the use of more traditional significance levels, and it was not until this mode was adopted that reasonably good results were obtained.

References

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

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

Cliff, N. Test theory without true scores? *Psychometrika,* in press.

Cliff, N., Cudeck, R., & McCormick, D. J. Evaluations of implied orders as a basis for tailored testing using simulations (Technical Report No. 4). Los Angeles: University of Southern California, Department of Psychology, September 1977.

Cudeck, R., Cliff, N., & Kehoe, J. TAILOR: A FORTRAN program for interactive tailored testing. *Educational and Psychological Measurement,* 1977, *37,* 767-769.

Cudeck, R., McCormick, D. J., & Cliff, N., Monte carlo evaluation of implied orders as a basis for tailored testing. *Applied Psychological Measurement,* 1979, *3,* 65-74.

Jensema, C. J. The validity of Bayesian tailored testing. *Educational and Psychological Measurement,* 1974, *34,* 757-766.

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

Loevinger, J. The attenuation paradox in test theory. *Psychological Bulletin,* 1954, *51,* 493-504.

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.

Lord. F. M. A theoretical study of two-stage testing. *Psychometrika,* 1971, *36,* 227-241.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley, 1968.

McBride, J. R. Some properties of a Bayesian adaptive ability testing strategy. *Applied Psychological Measurement,* 1977, *1,* 121-140.

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

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

Urry, V. W. Approximations to item parameters of mental test models and their uses. *Educational and Psychological Measurement,* 1974, *34,* 253-269.

Urry, V. W. Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement,* 1977, *14,* 181-196.

Weiss, D. J. *Final report: Computerized ability testing, 1972-1975.* Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1976.

Weiss, D. J. (Ed.). *Proceedings of the 1977 computerized adaptive testing conference.* Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1978.

Acknowledgments

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

Author's Address

Send requests for reprints or further information to Norman Cliff, Department of Psychology, University Park, University of Southern California, Los Angeles, CA 90007.