Implied Orders Tailored Testing: Simulation with the Stanford-Binet

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

Tailored testing by Cliff's (1975) method of implied orders was simulated through the use of responses gathered during conventional administration of the Stanford-Binet intelligence test. Tailoring eliminated approximately half the responses with only modest decreases in score reliability. Responses in tailored tests were shown by the Spearman-Brown prophecy formula to be equivalent to from 1.09 to 1.48 conventional items. Ninety-five percent of all responses implied by the tailoring procedures were identical to responses actually obtained during live testing.

Cliff (1975) proposed a method for tailored/adaptive testing distinct from up-and-down branching models (e.g., Lord, 1971) and traceline estimation procedures (e.g., McBride, 1977; Urry, 1977). Cliff suggested that the ordering information provided by the test responses and by the transitivity of such relations could be used to construct a joint order of persons and items and that this relationship could be exploited for tailored testing purposes.

In the simplest case, if Person A correctly answers Item B, which is in turn missed by Person C, a joint ability-difficulty order is evident, A>B>C. In addition to the two observed relations, A>B and B>C, there exists an implied relation A>C, which is the logical result of the relations that were actually observed.

In more complicated instances, relations can be connected in long chains to imply order between persons and items which these particular persons have not been given. The process for obtaining these chains, for adjusting them for the presence of inconsistency, and for assessing their statistical significance has been described in detail; technical descriptions of the rationale for this method and of the computer programs that implement it are given in Cliff (1975), Cudeck, Cliff, and Kehoe (1977), McCormick and Cliff (1977), and McCormick (1978). A more general review of the system is presented in Cliff, Cudeck, and McCormick (1979). The Cudeck et al. (1977) version of the implied orders system, called TAILOR, was used in the current study. This program was designed for the simultaneous administration of tailored tests to groups of examinees. An individual testing program has also been written (McCormick, 1978; McCormick & Cliff, 1977).

TAILOR is applied to a set of items on which there is no pretest information. Initial item-person assignments are made at random. When a few responses are available, TAILOR forms a tentative partial order of items and of persons, using these to guide the assignment of later items to persons. As new responses become available, the orders are updated and refined. At

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 4, No. 2 Spring 1980 pp. 157–163

any given time, it uses the current order of item difficulty of the items and each individual's responses to date to attempt to estimate his/her responses to the items that have not yet been taken. The process is complete when each person has a full complement of obtained or implied responses. The individual's final score can be the simple sum of responses that are either correct or predicted to be correct. However, as the procedure is actually implemented, a "net dominance score," which is highly correlated with number correct, is used but is slightly modified to take account of the person's standing relative to the other persons taking the test.

Previous Evaluations

Two previous evaluations of this testing method have been reported—one of which involved errorless data of a Guttman scale type (Cudeck, McCormick, & Cliff, 1979) and a second which used artificial responses generated according to the Birnbaum (1968) model (Cliff et al., 1979). Both studies found that scores obtained from a tailored test containing one-third to one-half the total number of items had only slightly lower test-retest reliability than complete test scores.

In the study using errorless data (Cudeck, McCormick, & Cliff, 1979) the rank correlation of tailored scores with true scores was limited only by the existence of person pairs that could not be untied by an item of intermediate difficulty. This typically occurred because no appropriate item existed. The average tau between tailored scores and assigned true scores was .96; and to attain this level of accuracy, only 48% of the possible items were presented.

In the study that used data generated according to the three-parameter Birnbaum (1968) model (Cliff et al., 1979), an average product-moment correlation of .89 was obtained between tailored scores and assigned true scores (tau=.76). The Pearson correlation from conventional complete test matrices drawn under the same variety of parameter values was .93 (tau=.81). Thus, the tailored tests were nearly as reliable as complete tests, although based on substantially reduced numbers of items.

From the evidence gathered to date, this method for tailored testing appears capable of introducing substantial efficiencies into the process of administering tests. To bring the evaluation procedures one step closer to live testing, and to accumulate further experience with different data types, the current simulation was undertaken using a large file of responses obtained from actual administrations of the Stanford-Binet intelligence test.

Method

Nature of the Data

The data source was a file of responses of 622 children to the 122 items of the Stanford-Binet. The ages of the children were roughly uniformly distributed from 24 months to 178 months. The mean IQ was 117.3, with a standard deviation of 17.6, and ranged from 66 to 166. Thus, the sample was well above average but quite variable in IQ.

Because the apparent reliability and item discrimination resulting from such a wide range of ages would be unreasonably high, six groups of children were determined according to their ages to form more homogeneous groups. The characteristics of the six groups are shown in Table 1.

Test item discriminations were defined with reference to the variance of the group tested. Therefore, average item discrimination could be indirectly manipulated by influencing total score variance. This was done by selecting two types of groups—Wide age range and Narrow age range—on the basis of chronological age.

First, the three Wide age-range groups were selected. Ages in these groups ranged from 24 to 59 months, from 60 to 95 months, and from 132 to 179 months. Appropriate item pools were selected from the 122 items of the complete test by deleting those that were either missed by all children in a group or were answered correctly by

Table 1
Characteristics of Age groups

Age (years)	N	Binet Levels	Number of Items	Mean Discrimination $\bar{x}_a$	Difficulty Mean $\bar{x}_b$	Difficulty S.D. sd_b
Wide						
2-4	156	2-6 to 8-0	54[a]	0.70	.095	1.84
5-7	179	4-0 to 14-0	72[b]	1.71	-.287	1.84
11-14	130	7-0 to Ad.3	74[a]	1.26	.562	1.59
Narrow						
3	65	2-6 to 8-0	54			
6	62	4-0 to 14-0	72			
13-14	60	7-0 to Ad.3	74			

[a]Include 2 items with zero variance, not used to calculate mean or standard deviation
[b]Include 1 item with zero variance, not used to calculate mean or standard deviation.

all. For the remaining items, difficulty and discrimination indices were computed (Urry, 1974) based on the observed proportion correct and item biserial correlations within each group. These statistics, as well as the size and range of the final item pools, are also shown in Table 1.

The three Narrow age-range groups, also in Table 1, represented even more restricted age ranges. The first two groups consisted of all children within single years. For the oldest level it was necessary to include two different years rather than one, simply to obtain an adequately large pool. The three Narrow age-range groups were subsets of the Wide age-range groups, but the item pools were kept the same in both instances. Because of the small size of the Narrow age-range groups, item statistics were not recomputed for these subsamples. It seems reasonable to assume, however, that item discriminations would have been reduced along with the range of ages.

Procedure

Within each age group, random samples of persons were taken. For each sample of n persons, two separate random samples of 25 items were selected. Thus, each sample of data consisted of two $n \times 25$ matrices of item scores. For reference, the two were called Form T and Form C. Conventional total scores on each of the forms were derived, hereafter called Complete scores. Form T was designated as the basis for the operation of TAILOR. The system utilized this matrix much as it would interact with human subjects except that instead of scoring a response from a real subject, the program simply examined the appropriate element of the response matrix for Form T to determine whether it was correct. In this way a Tailored score was derived for each person in the sample. Thus, there were two data matrices of size $n \times 25$, designated T and C. Three sets of scores were obtained: Complete scores from each of T and C and a Tailored score from T.

The major variable used to evaluate TAILOR has been its correlation with an independently derived score. In the case of the data model simulations (Cliff et al., 1979), this was the correlation of the Tailored score on a set of simulated items with the true scores that were used to generate the data. In the present instances with real

data, there was no true score. Instead, the correlation of Tailored scores with the Complete scores obtained on the parallel items (Form C) was used. These formed two experimentally independent ability estimates. This correlation was compared to the parallel forms reliability obtained by taking the Complete scores of both Forms T and C. For convenience, the Tailored score/Complete score correlation also will be referred to later in this paper as a reliability correlation. The critical comparison was then between the reliability of the Tailored scores of Form T and Complete scores of Form C with the reliability of both Complete scores.

Five random samples of 20 persons were taken within each age group to furnish replications for the simulation. The size of the populations available for sampling meant that there was some overlap between the samples. This did not have a spurious effect on the correlation between the two scores, since they were always based on separate items; but it did mean that the five replications were likely to be somewhat more similar to each other than they would be if taken from an infinitely large population. This effect was small, however, particularly since items were also sampled, and it was essentially irrelevant to the present purposes.

To summarize, for each replication within each age range, there were conventional test scores on two randomly parallel forms. These scores were correlated to give Complete-Complete (C-C) correlations. One of the forms was also the basis for a Tailored score, and the correlation of this with the Complete score on the other form was also computed (C-T correlations). These correlations, both Pearson and tau, were the major dependent variables in this study. There were three Wide age-range groups and three Narrow age-range groups, two sample sizes (20 and 40), and five replications within each, making a total of 60 applications of TAILOR for this study.

Results

Table 2 presents the principal results of the simulated testing. Included in the table are the average proportion of responses required of each subject in each of the tailored test conditions. TAILOR completed the score matrix on the basis of responses to about half the 25 items, slightly more if there were 20 persons, slightly less if there were 40.

Table 2 also shows the average C-T and C-C correlations for each condition. For the Wide age-range groups the overall average reliability estimates using tau were .71 and .75 for C-T and C-C cases, respectively. For the Narrow age-range groups the corresponding taus were .62 and .68.

Table 2
Complete and Tailored Parallel Form Correlations (tau) and Proportion of Items used by TAILOR

Dependent measure	N	Wide Age Range				Narrow Age Range			
		2-4	5-7	11-14	$\bar{x}$	3	6	13-14	$\bar{x}$
C-C	20	.72	.74	.74	.73	.65	.65	.73	.68
C-T	20	.69	.74	.70	.71	.57	.62	.67	.62
Prop. of Items	20	.54	.56	.56	.55	.54	.56	.56	.56
C-C	40	.76	.76	.77	.76	.64	.66	.73	.68
C-T	40	.73	.71	.72	.72	.58	.60	.67	.62
Prop. of Items	40	.44	.45	.47	.46	.46	.49	.49	.48

To determine the relationship between C-C reliability and C-T reliability a simple regression was performed using the mean correlations from Table 2. A Pearson correlation of .95 was found between the 60 C-T and C-C reliabilities. The regression equation for predicting C-T from C-C yielded a slope of 1.17, indicating that the reliability of the Tailored score changed somewhat more than that of the Complete score as the latter varied. This oversensitivity of the tailored tests to the reliability of the complete tests is not unique to the data analyzed here but was also found in the monte carlo study reported by Cliff et al. (1979).

At the conclusion of each test session, TAILOR provided a predicted Complete score matrix, composed in part of the person's actual responses, which were used to imply his/her other responses, and in part of the responses that were deduced by TAILOR from them. Since this was a simulation, the person's actual responses to the latter were also available and could be compared to the predicted item responses. Table 3 shows the proportions of correctly predicted responses for each of the various conditions. More than 95% of the predictions made by TAILOR were correct.

Relative Efficiency of Tailoring

One way of looking at the process of tailored testing is to consider it an item-culling procedure, removing those items from a test that are least useful for testing a particular examinee. A natural way to evaluate the results of such test shortening is to compare the tailored test, which is shortened by design, to a test shortened by random elimination of items. The assessment of such a comparison can be made by means of an adaptation of the Spearman-Brown prophecy formula,

$$k = \frac{r'^2 - r'^2 r}{r^2 - r'^2 r} \qquad [1]$$

where

- r is the correlation of two parallel forms of equal length,
- r' is the correlation of one test of this length with a second of altered length, and
- k is the proportion of increment or decrement to the original length.

The formula is not the same as the more familiar one because here concern was with the estimated correlation of a shortened test with a full-length parallel form.

In this context the correlations between C-T and C-C correspond to r' and r in Equation 1 and can be used to solve for the length of a randomly shortened test that would have the same correlation with the complete test as was shown by the tailored test. Since the proportion of items used by TAILOR to obtain the same correlation was known, a ratio of the proportion used in a random selection, k, divided by the proportion used by TAILOR shows the number of items from a randomly shortened test necessary to do the work of a single item from a tailored test.

Table 4 shows the various measures required for calculating such a ratio and the computed values for the four different types of data. As can be seen, the efficiency ratio ranged from a high of 1.48, for the 40 persons in the Wide age-range groups, to a low of 1.09, for 20 persons in the

Table 3
Proportion of Correctly Predicted Responses

Number of Persons	Wide Age Range				Narrow Age Range			
	2-4	5-7	11-14	$\bar{x}$	3	6	13-14	$\bar{x}$
20	.95	.98	.96	.96	.93	.98	.94	.95
40	.96	.96	.95	.96	.95	.97	.95	.96

Table 4
Efficiency Calculations

Statistic	Wide Age Range		Narrow Age Range	
	20 Persons	40 Persons	20 Persons	40 Persons
Percent Responses	.55	.46	.56	.48
C-T r	.83	.87	.77	.75
C-C r	.86	.89	.81	.81
k	.70	.67	.60	.56
Ratio k/percent	1.26	1.48	1.09	1.17

Narrow age-range groups. It has already been shown that the reliability of Tailored scores increased disproportionately with the reliability as indexed by Complete test scores (Cliff et al., 1979). This may account for the increase in efficiency from the Narrow to Wide age-range data observed here.

Increases in tailored test efficiency also occurred as the number of examinees increased. Because persons are sources of information about item order, the more persons taking the test, the fewer responses are required from each to establish a reliable item order. Therefore, because the item order is determined earlier in the testing session, the items presented to each examinee are more quickly located at the examinee's level of ability and the testing is more efficient. This effect is not large, however.

Discussion

Results from the current simulation agreed in several ways with previous evaluations. Although the dependent measures used in the previous studies were the correlations of Complete and Tailored scores with true scores, rather than with a parallel set of Complete test scores as they were here, the relationship between measures evaluating tailored tests and those representing complete test performance was similar. In spite of a substantial shortening of the test, the reliability of Tailored scores, measured either as a correlation with true scores or as a correlation with a parallel complete test, was only slightly less than the reliability of Complete scores. The extent of test shortening observed in the current simulation was, on the average, at the lower end of the one-half to two-thirds items eliminated in previous studies but was not outside the range. In addition, as in the second evaluation study (Cliff et al., 1979), the reliability of Tailored scores was found to be highly correlated with the reliability of Complete test scores.

Also observed in both the current and Birnbaum model studies was a tendency for Tailored scores to fall off in quality faster than Complete test scores as a function of the overall quality of the test data. Finally, in all three evaluation studies, TAILOR produced better results when the number of persons was increased. As long as the data are consistent, the number of person-item relations needed to order a set of items does not change with the number of persons who take the test. Therefore, when more people take the test, fewer relations must be contributed by each person to obtain a given precision of ordering. Because the items can be ordered after fewer responses per person, the accurate matching of persons and items comes earlier in the test. Thus, the persons are also ordered more efficiently. Although there was an effect of sample size, perhaps more striking is the absolute sample size used here. Effective tailoring took place with groups as small as 20, and this is in the absence of any use of pretest information concerning item difficulties.

In addition to extending old findings with a new type of data, two new methods of evaluating

tailored test data were implemented. The first compared the item relations predicted by TAILOR with those known to be the outcomes when the items were actually given. The high proportion of correct predictions (.93 to .98) is evidence of the accuracy of the process. That is, when TAILOR gathers about half the responses, it can predict the remainder with about .95 accuracy.

The second method was to compute the ratio of the number of items needed by TAILOR to attain a given correlation with the parallel form Complete test score and the number of items predicted by an adaptation of the Spearman-Brown formula to be required if items were selected randomly. The last measure can be computed for any tailoring system, and no doubt the values reported here would become more useful if other tailoring methods were similarly evaluated.

Conclusions

The results of several analyses of data from simulated tailored testing indicated that the TAILOR procedure could reduce the length of a test substantially with only modest decreases in score reliability. It should be emphasized that the data set for this analysis was based on a thoroughly researched item pool with exceptional characteristics, so no unqualified generalization to other tests can be made. However, the investigations to date appear to support the conclusion that this model is appropriate for many situations. The next crucial extension is to assess the procedure in a live tailored testing administration. A report on such a study is currently in preparation.

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., Cudeck, R., & McCormick, D. J. Evaluation of implied orders as a basis for tailored testing with simulation data. *Applied Psychological Measurement*, 1979, *3*, 495-514.

Cudeck, R., Cliff, N., & Kehoe, J. TAILOR: A FORTRAN procedure 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.

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

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

McCormick, D. J. TAILOR-APL: An interactive computer program for individual tailored testing. (Technical Report 5). Los Angeles, CA: University of Southern California, Department of Psychology, 1978.

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

Urry, V. W. Approximations to item parameters of mental test models and their uses. *Educational and Psychological Measurement*, 1974, *37*, 253-269.

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

Acknowledgments

This research was supported by the Office of Naval Research, Contract N00014-75-C-0684, NR150-373. The data were made available by Mark Reckase of the University of Missouri. They were collected under the direction of Byron Egeland, now at the University of Minnesota, and by Bill Curlett and John Plew at Syracuse University.

Author's Address

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