AD-A053 103 UNIVERSITY OF SOUTHERN CALIFORNIA LOS ANGELES DEPT O--ETC F/G 5/10
TAILOR-APL: AN INTERACTIVE COMPUTER PROGRAM FOR INDIVIDUAL TAIL--ETC(U)
MAR 78 D J MCCORMICK N00014-75-C-0684
UNCLASSIFIED TR-5 NL

| OF |
AD
A053103

END
DATE
FILMED
5-78
DDC

(12)

TAILOR-APL: An Interactive Computer Program for Individual Tailored Testing

Douglas McCormick

Technical Report No. 5

Department of Psychology

University of Southern California

Los Angeles, California 90007

March, 1978

Prepared under contract No. N00014-75-C-0684
NR No. 150-373, with the Personnel and
Training Research Programs, Psychological Sciences Division

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: Technical Report No. 5 (14) TR-5	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): (6) TAILOR-APL: An Interactive Computer Program for Individual Tailored Testing	5. TYPE OF REPORT & PERIOD COVERED: (9) Technical rept.
7. AUTHOR(s): (10) Douglas J. McCormick	8. CONTRACT OR GRANT NUMBER(s): (15) N00014-75-C-0684
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Psychology, University of Southern California, Los Angeles, California 90007	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 61153N, RR042-04, RR042-04-01, NR 150-373
11. CONTROLLING OFFICE NAME AND ADDRESS: Personnel and Training Research Program, Office of Naval Research (Code 458), Arlington, Virginia 22217	(11) 12. REPORT DATE: Mar 78 (12) 67 pp. 13. NUMBER OF PAGES: 64
(16) 14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): RR042 04 (17) RR042 04 01	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release, distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

DDC RECEIVED APR 24 1978 F

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

(U) Testing, (U) Computer, (U) Tailored, (U) Adaptive, (U) Response Contingent

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

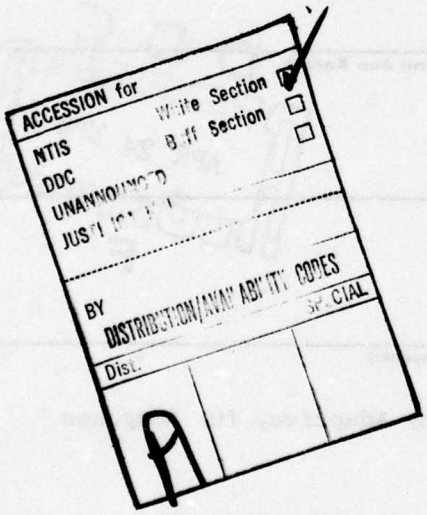
A computer program for administering individual tailored tests was evaluated with 50 examinees and compared with a complete test on the criterion of odd-even reliability. Although the tailored tests presented only 0.44 as many of the items as the complete tests, the reliability of the scores was 0.83 compared to a complete test reliability of 0.78. In —>

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601 |

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

400 762

addition, these results were obtained without pretesting of items.



TAILOR-APL: An Interactive Computer Program for Individual Tailored Testing

Anyone can create a test simply by gathering together a set of questions. A good test will have some relevance to an ability or quantity of interest and produce reliable scores for the full range of examinees for which it is intended. Tailored testing methods are individualized testing methods which make use of the fact that the items which reliably measure an attribute of a single individual need only be a small subset of the items necessary for measuring a group.

Rudimentary tailoring began with the Binet Intelligence Test and exists in other individual tests such as the Wechsler Intelligence Scales (Wechsler, 1958) and the Peabody Picture Vocabulary Test (Dunn, 1965). These tests begin at levels of difficulty, estimated from age and other indicators, where a specified number of items in a sequence or within an age level will all be answered correctly and continue to more difficult items until a long string of errors marks the probable limit of success. It is assumed that items of lower difficulty than those administered would have been answered correctly and items of greater difficulty than those administered would have been incorrect. Whether extending the range of the Stanford-Binet would produce different scores was investigated by Bradway (1943) who found no significant difference in absolute socres or score reliability.

The first systematic attempt at tailoring came in 1946 in a paper by Cowden. Cowden applied Wald's (1947) sequential analysis techniques, originally used for industrial product testing, to the special problem of tailored testing when only two outcomes are possible (pass-fail, accept-reject,

hire-don't hire, etc.). Items which produce the greatest differentiation between groups are presented in the same order for each examinee. As soon as the responses from a particular individual allow him to be classed with a specified certainty in either group his test is ended. Sequential analysis has been most recently applied to criterion referenced testing in computer assisted instruction programs (Wood, 1970; Ferguson, 1969; Ferguson & Hsu, 1971).

Hick in 1951 suggested a rationale for tailored testing using ideas from his work in signal detection and information theory. His notion was that the item providing the maximum amount of information was that item which an individual has a .50 probability of answering correctly. Consequently, the initial item of a tailored test should be the item of mean difficulty in the individual's population. If the first question is answered correctly, the second item should be answered correctly by .50 of the people who answered the first item correctly and so on.

Perhaps the farthest strategy from Hick's ideal, but the easiest tailoring system to administer without a computer is the "two stage" test (Angoff & Huddleston, 1958; Cleary, Linn & Rock, 1968a, 1968b; Linn, Rock & Cleary, 1969; Lord, 1971a). The two stage test has an initial test, often much shorter than the second, which routes all the examinees according to their scores to a final test appropriate to their ability levels. A variety of such compromise strategies exist between the systems which branch after every item and a conventional test. The number of stages and items available at each stage vary as do the scoring and branching rules. Lord alone investigated 200 approaches to two stage testing.

The next type of system has been the most prolific (for a large list of

sources and an excellent overview of tailored testing see Weiss & Betz, 1973). This approach structures the item pool into a decision tree that looks similar to Pascal's triangle. Instead of providing two unique items for each branching, the branches form a lattice of reconverging paths. A person whose answers are r-r-w-w-w ends up in the same place as a person whose answers are w-w-w-r-r or any other combination of as many right and wrong answers. Instead of $2^n - 1$ items which would be required for the structured item pool of a test in which n items are actually presented, if each branch were unique, the reconverging structure requires $n(n + 1)/2$ items. Both totals are excessive for all but very short tests. There are other problems with the structured item pool such as fitting the size of the branching steps up and down (or rather from side to side) within the order of difficulty to conform with Hick's prescription. To get the proper conditional probabilities of a correct response at each juncture, a shrinking step size is necessary which is incompatible with the equal size units of the reconverging triangle.

One researcher (Mussio, 1972) attempted to reduce the inordinate item pool requirements by truncating the lower corners of the triangle, but the most satisfactory solution to the many problems of a structured item pool is to unstructure it.

An unstructured item pool is primarily one dimensional, with the possibility of item discrimination being used as a second dimension. Items are chosen at each stage in a tailored test according to their actual difficulty, which may not be the ideal .50, but will be as close to it as possible.

Lord (1971b) implements an unstructured method which he calls the "flexilevel" test. Beginning with the item of expected .50 difficulty, the examinee branches up or down one item for each correct or incorrect response. When an individual is forced to double back and confronts items which they have

already taken they are simply skipped over. The test is ended when half the items have been taken. The result of this technique is that the examinee takes the half of the item pool closest to their ability level. Although far from ideal, it is a very easy test to administer and the individual's score is simply the number correct.

Among the unstructured approaches are three methods that use extensive calculations after each response to determine the best item available for presentation.

Novick (1969) has suggested a possible Bayesian method of tailored testing. Beginning with a population distribution as the initial prior, the appropriate impact of a correct or incorrect answer to each question is seen in the posterior distributions as a narrowing of the variance and a movement of the mean to a higher value for correct responses and to a lower value for incorrect responses. Each item is chosen to give the maximum reduction of variance. The posterior distribution from each item becomes the prior distribution for the next. The process is continued until the posterior variance is less than a predetermined maximum.

Owen (1969, 1970) has produced a Bayesian algorithm for actual implementation which involves a simplifying assumption of normal priors. In addition, Owen has incorporated a method of dealing with the possibility of the correct answer in a multiple choice format being guessed.

Urry (1970) proposed a maximum-likelihood method of doing very much the same thing as the Bayesian procedures do, but without a prior distribution. Rather than assuming a flat prior, or a population prior as the Bayesian methods do, the maximum-likelihood methods establish an initial probability distribution on the basis of one correct and one incorrect response. In

order to do this items are presented at the beginning of the procedure that are at extreme positions in the difficulty continuum. Such items have very low information value and represent therefore an inefficiency in maximum-likelihood methods.

In the case of maximum-likelihood approaches an operational system was provided by Reckase (1974). As in the implementation of Bayesian techniques, simplifying assumptions were made to reduce the complexity of necessary computations. The Reckase procedure is based on the Rasch model (Rasch, 1960) which treats all items as if they had equal discrimination and makes no allowance for guessing.

All of the preceeding methods of tailored test administration begin with knowledge of the difficulty levels of the various items based on previous conventional testing. The more elegant methods also require calculation of item discrimination and guessing probabilities. The accuracy of the results depends on the extensiveness of pretesting. Gugel, Schmidt, and Urry (1976) analyze the results obtained from Owen's method using a range of from 500 to 2,000 pretest examinees.

Applying these methods to well established tests which have already been extensively pretested would not be difficult, but adding items would be a slow process. For the majority of tests which are not maintained for thousands of examinations, the effort of pretesting is likely to be prohibitive. There is also the possibility that the sheer volume of pretesting would encourage the use of item parameters estimated from the responses of an inappropriate sample.

Implied Orders

The research to be reported in this paper was undertaken to produce and

evaluate by live testing a test tailoring mechanism described by Cliff (1975) called TAILOR in its group computer program form (Cudeck, Cliff & Kehoe, 1977) and TAILOR-APL in the individual testing form (McCormick & Cliff, 1977) which is the version evaluated here.

As outlined by Cliff (1975), the origins of TAILOR are in ordinal scaling, and its approach to test tailoring emphasizes the order relations among persons and items. In addition to an emphasis on order relations, TAILOR presents a unique solution to the problem of gathering item information. The program begins with no knowledge of item difficulties or other item characteristics and makes the collection of item information part of test administration. The efficiency of tailoring at any time is therefore determined by the thoroughness of the information so far collected. In this way significant tailoring can be enjoyed by examinees who would have been forced by the pretesting requirements of all other tailoring programs to take complete tests.

The series of matrix operations which define TAILOR take place in the context of an expanded person-item binary score matrix. This is depicted in Figure 1. Instead of a conventional persons x items matrix in which the non-zero entries represent successes of persons, characterized as rows, with items, represented by columns, the persons + items x persons + items matrix represents four types of relations. The intersection of a person's row with an item's column can mean what it did before, but can also represent an answer that was implied to be correct based on the individual's previous answers rather than being an actual response to an administered item. The intersection of an item's row with a person's column, if non-zero, represents either a wrong answer or an answer which was implied to be wrong. Person-person and item-item intersec-

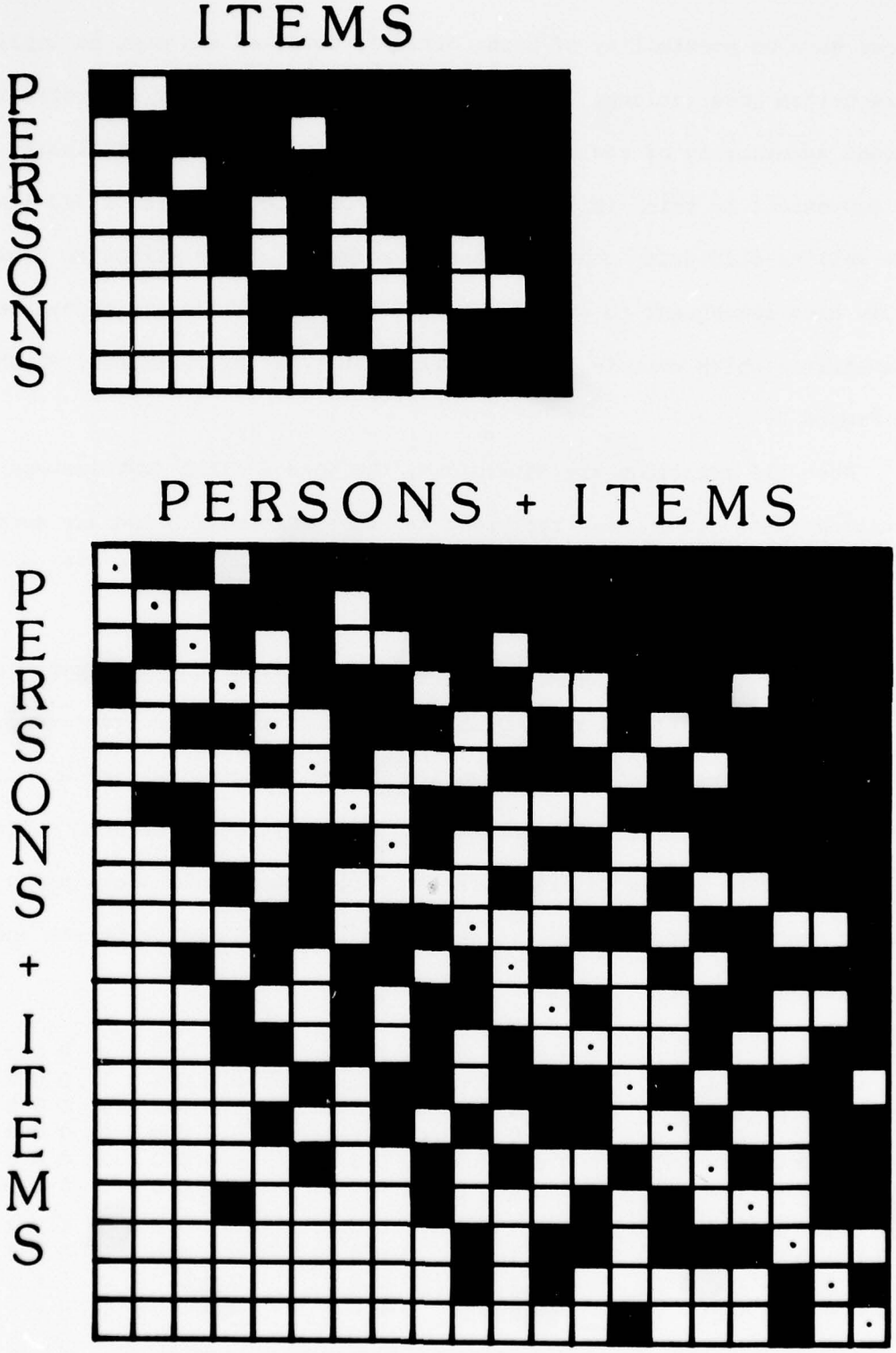


Figure 1

tions have no possibility of being directly observed and must be implied from person-item observations. Person-person and item-item entries refer to significant superiority of row over column either in ability or difficulty. It is convenient to think in terms of a joint ordering of persons and items on the ability-difficulty continuum and to refer to all relations as dominances. It is also convenient to order items and persons arbitrarily to create four submatrices which contain the four different types of relations, as shown in Figure 2.

When all relations are determined, the person - item and item-person matrices (wins and losses) represent the same information and are matrix transpose-complements of each other.

Deriving Item-Item and Person-Person Dominances from Observed Responses

We begin a tailored test by observing correct and incorrect responses to person-item pairs. This information can be recorded as ones and zeroes in the expanded binary score matrix, A. Multiplying the persons + items x persons + items matrix by itself is the first step in the implication process. This produces a matrix, A^2, with entries only in the person-person and item-item intersections:

$$\begin{bmatrix} 0&0&0&0&0&0&1&1&0&1 \\ 0&0&0&0&0&0&0&1&1&1 \\ 0&0&0&0&0&0&0&1&0&1 \\ 0&0&0&0&0&0&0&0&1&1 \\ 0&0&0&0&0&0&0&0&0&1 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&1&1&1&1&1&0&0&0&0 \\ 0&0&0&1&0&1&0&0&0&0 \\ 0&0&0&0&1&1&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0&0 \end{bmatrix} \begin{bmatrix} 0&0&0&0&0&0&1&1&0&1 \\ 0&0&0&0&0&0&0&1&1&1 \\ 0&0&0&0&0&0&0&1&0&1 \\ 0&0&0&0&0&0&0&0&1&1 \\ 0&0&0&0&0&0&0&0&0&1 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&1&1&1&1&1&0&0&0&0 \\ 0&0&0&1&0&1&0&0&0&0 \\ 0&0&0&1&1&1&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0&0 \end{bmatrix} = \begin{bmatrix} 0&1&1&2&1&3&0&0&0&0 \\ 0&0&0&2&1&3&0&0&0&0 \\ 0&0&0&1&0&2&0&0&0&0 \\ 0&0&0&0&1&2&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&2&2&4 \\ 0&0&0&0&0&0&0&0&1&1 \\ 0&0&0&0&0&0&0&0&0&2 \\ 0&0&0&0&0&0&0&0&0&0 \end{bmatrix}$$

$$A \qquad A \qquad = \qquad A^2$$

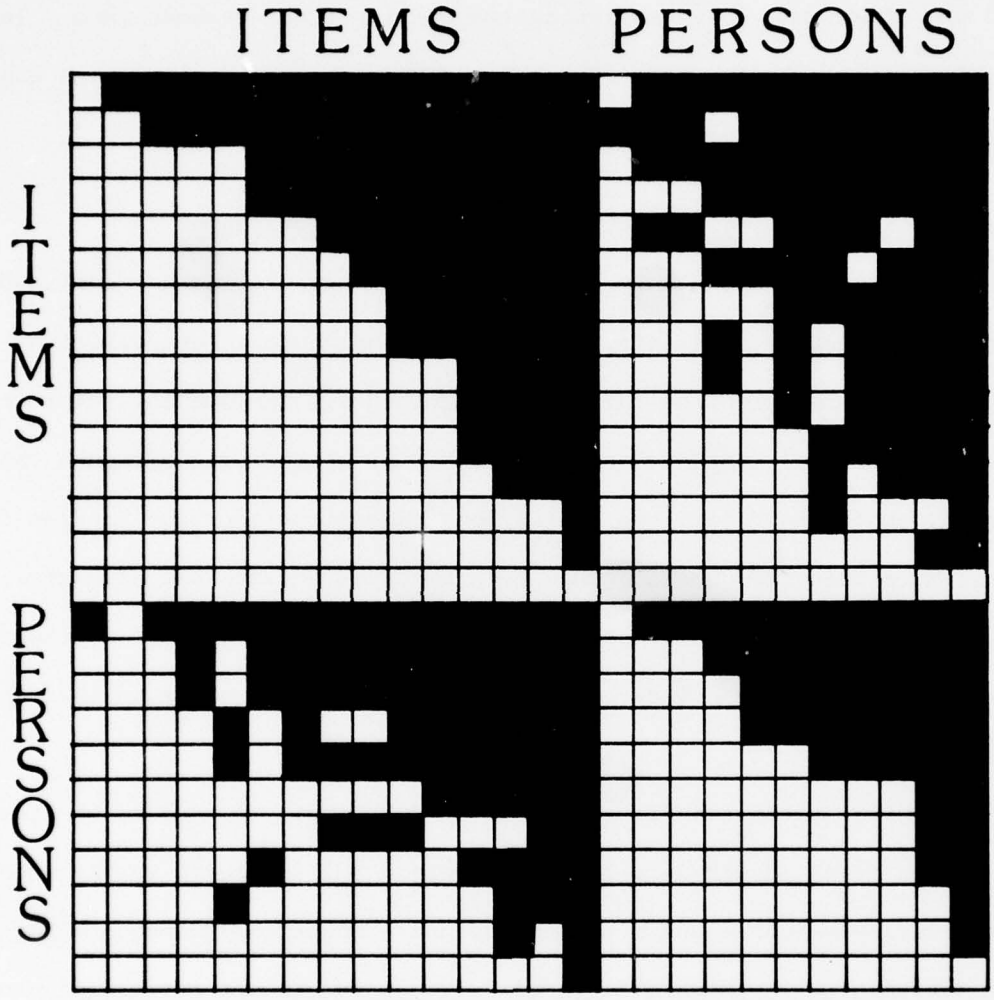


Figure 2

The entry in a person-person intersection represents the number of items which were answered correctly by the row person and answered incorrectly by the column person. Items which both missed or both answered correctly produce no entries and are of no use in establishing the ability order between them. In a similar fashion entries in the item-item submatrix represent persons which are dominated by the row item and dominate the column item.

Testing Corresponding Entries for "Significant" Dominance

To establish the binary order relations once the integer dominance matrix has been computed, each entry representing dominance in one direction is compared to the entry representing the reverse dominance. The statistical rules for deciding which element dominates the other or whether no dominance can be established, are divided into two approaches. The first approach handles cases where more than one dominance has been observed between two elements. The second approach is designed specifically for the instance where a single dominance has been recorded in one direction and no counter-dominance in the opposite direction.

Looking at the relationship between two people, the number of items missed by person i and answered correctly by person j is compared to the number answered correctly by person i and missed by person j according to McNemar's formula for determining the significance of differences between correlated proportions (Guilford & Fruchter, 1973, p. 165).

McNemar's Test

Based on items which both persons have taken.

Person i \ Person j	won	lost
won	A	B
lost	C	D

$$Z_{(i>j)} = \frac{B - C}{\sqrt{B + C}}$$

If the Z statistic produced by McNemar's test exceeds 1.0, "significant" dominance is recorded by entering a one at the appropriate intersection of a binary person-person submatrix of the persons + items x persons + items matrix.

In Monte Carlo investigations of simulated testing which used a group testing Fortran version of TAILOR, it was found that the discrete jump from no implications to the case where one dominance is observed in a single direction did not allow adequate precision in establishing significance. If the one-zero case were allowed, too many false implications flooded the matrix. If the one-zero case was not allowed very little tailoring occurred. An intermediate criterion, a second significance test, was developed on the basis of binomial probabilities specifically to handle one-zero cases. The equation for this criterion is:

$$P = \frac{\binom{N - I}{J}}{\binom{N}{J}}$$

N = number of items

I = person i's total items correct

J = person j's total items wrong

The significance test is based on the row of S (the correct answer or wins matrix) which contains all of the successful individual's wins including the current item. Also used for the test is the column of $\tilde{S}'$ which contains the losses of the unsuccessful person. In the instance of a one-zero case, one of the wins in the dominant person's row corresponds to a loss in the dominated individual's column. In order to determine the significance of this correspondence, the binomial probability is calculated for the event that the wins in the row and the losses in the column would form no correspondence if randomly distributed. If the probability of no correspondence is higher than .5 the dominance relation is retained as a one in the binary version of the person-person submatrix, $\overline{A^2}$ (binary).

In the first row of A^2 (shown below) there are two instances of total dominances greater than one. There are two dominances of item one over item four and three of item one over item six. Both cases are handled by McNemar's test and since no counter-dominances exist, both were maintained in the binary version of A^2, which is $\overline{A^2}$. (The symbol $\triangleq$ shall represent the conversion of integer entries to binary relations.) Also in the first row are three instances where a single dominance exists of item one over another item. Again, no counter-dominances exist. Because only a single dominance is involved, these cases are handled by the binomial probability procedure. The probability that the first two might have occurred by chance is less than .5 so they are retained in $\overline{A^2}$. The dominance of item one over item five has better than a .5 probability of occurring simply by a random assortment of the persons who missed item one and those who answered five correctly, so it is not retained.

$$\begin{bmatrix} 0&1&1&2&1&3&0&0&0&0 \\ 0&0&0&2&1&3&0&0&0&0 \\ 0&0&0&1&0&2&0&0&0&0 \\ 0&0&0&0&1&2&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&2&2&4 \\ 0&0&0&0&0&0&0&0&1&1 \\ 0&0&0&0&0&0&0&0&0&2 \\ 0&0&0&0&0&0&0&0&0&0 \end{bmatrix} \cong \begin{bmatrix} 0&1&1&1&0&1&0&0&0&0 \\ 0&0&0&1&0&1&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&1&1&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&1&1&1 \\ 0&0&0&0&0&0&0&0&1&0 \\ 0&0&0&0&0&0&0&0&0&1 \\ 0&0&0&0&0&0&0&0&0&0 \end{bmatrix}$$

$$A^2 \qquad \cong \qquad \overline{A^2}$$

The effect of the binomial probability test of significant dominance is to limit implications in the one-zero case to the earlier stages of testing when information is scarce. As the vectors fill, the probability of a random one-zero correspondence increases and stronger evidence is required.

The procedure for determining significant order relations between two items proceeds analogously. Integer products of the first multiplication of the persons + items matrix by itself are tested for significant dominance whether they are item-item dominances or person-person dominances. Item-item entries are the result of persons who were dominated by the row item and in turn dominated the column item.

Determining Higher Order Relations

After significance testing there are item-item and person-person relations recorded as binary entries. Analogous to the process of determining these relations by looking at items which were common to each pair of persons and persons which were common to each pair of items, item-item or person-person relations can be established by looking at items which share relationships with other items and persons which are shared in relationships with pairs of persons. For instance, if person A dominates persons B, C and D on the basis of items common to each one and A, and B, C and D all dominate person E, it

can be implied that person A dominates person E, assuming there are no such implications in the opposite direction. The new implications based on these person-person-person or item-item-item chains of implications are arrived at by significance testing of the integer products of the binary matrix containing person-person and item-item relations multiplied by itself, A^4. Repeated powering would produce entries representing longer and longer chains of

$$\begin{bmatrix} 0&1&1&1&0&1&0&0&0&0 \\ 0&0&0&1&0&1&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&1&1&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&1&1&1 \\ 0&0&0&0&0&0&0&0&1&0 \\ 0&0&0&0&0&0&0&0&0&1 \\ 0&0&0&0&0&0&0&0&0&0 \end{bmatrix} \begin{bmatrix} 0&1&1&1&0&1&0&0&0&0 \\ 0&0&0&1&0&1&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&1&1&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&1&1&1 \\ 0&0&0&0&0&0&0&0&1&0 \\ 0&0&0&0&0&0&0&0&0&1 \\ 0&0&0&0&0&0&0&0&0&0 \end{bmatrix} = \begin{bmatrix} 0&0&0&1&1&3&0&0&0&0 \\ 0&0&0&0&1&1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&1&1 \\ 0&0&0&0&0&0&0&0&0&1 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \end{bmatrix}$$

$$A^2 \qquad\qquad A^2 \qquad = \qquad A^4$$

$$\cong \begin{bmatrix} 0&0&0&0&1&1&0&0&0&0 \\ 0&0&0&0&1&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&1 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \end{bmatrix}$$

$$\cong \qquad \overline{A^4}$$

implication, but the empirical observation has been made that few useful implications are made beyond the first powering of the matrix. The new higher order relations are then combined with the relations of A^2 according to the rules of Boolean addition (denoted ●).

$$\begin{bmatrix} 0&1&1&1&0&1&0&0&0&0\\ 0&0&0&1&0&1&0&0&0&0\\ 0&0&0&0&0&1&0&0&0&0\\ 0&0&0&0&1&1&0&0&0&0\\ 0&0&0&0&0&1&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&1&1&1\\ 0&0&0&0&0&0&0&0&1&0\\ 0&0&0&0&0&0&0&0&0&1\\ 0&0&0&0&0&0&0&0&0&0 \end{bmatrix} \bullet \begin{bmatrix} 0&0&0&0&1&1&0&0&0&0\\ 0&0&0&0&1&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&1&0\\ 0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0 \end{bmatrix} = \begin{bmatrix} 0&1&1&1&1&1&0&0&0&0\\ 0&0&0&1&1&1&0&0&0&0\\ 0&0&0&0&0&1&0&0&0&0\\ 0&0&0&0&1&1&0&0&0&0\\ 0&0&0&0&0&1&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&1&1&1\\ 0&0&0&0&0&0&0&0&1&1\\ 0&0&0&0&0&0&0&0&0&1\\ 0&0&0&0&0&0&0&0&0&0 \end{bmatrix}$$

$$\overline{A^2} \bullet \overline{A^4} = \overline{A^2 \bullet A^4}$$

Implying Person-Item Relations from Observed Scores

Once person-person and item-item binary relations are established, implied person-item relations corresponding to right or wrong answers to test questions, can be determined through common items or common persons. The chains of implication are derived by multiplication of the original matrix of observed scores, A, by an expanded matrix, $\overline{(A^2 \bullet A^4)}$, containing only binary person-person and item-item relations. Each person-item and item-person integer entry is tested for significant dominance over its counterpart and the final relations preserved as binary entries.

$$\begin{bmatrix} 0&1&1&1&1&1&0&0&0&0\\ 0&0&0&1&1&1&0&0&0&0\\ 0&0&0&0&0&1&0&0&0&0\\ 0&0&0&0&1&1&0&0&0&0\\ 0&0&0&0&0&1&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&1&1&1\\ 0&0&0&0&0&0&0&0&1&1\\ 0&0&0&0&0&0&0&0&0&1\\ 0&0&0&0&0&0&0&0&0&0 \end{bmatrix} \begin{bmatrix} 0&0&0&0&0&0&1&1&0&1\\ 0&0&0&0&0&0&0&1&1&1\\ 0&0&0&0&0&0&0&1&0&1\\ 0&0&0&0&0&0&0&0&1&1\\ 0&0&0&0&0&0&0&0&0&1\\ 0&0&0&0&0&0&0&0&0&0\\ 0&1&1&1&1&1&0&0&0&0\\ 0&0&0&1&0&1&0&0&0&0\\ 0&0&0&0&1&1&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0 \end{bmatrix} = \begin{bmatrix} 0&0&0&0&0&0&0&2&2&4\\ 0&0&0&0&0&0&0&0&1&2\\ 0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&1\\ 0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0\\ 0&0&0&2&1&3&0&0&0&0\\ 0&0&0&1&1&2&0&0&0&0\\ 0&0&0&0&0&1&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0 \end{bmatrix}$$

$$\overline{(A^2 \bullet A^4)} \quad A = \overline{(A^2 \bullet A^4)}A$$

$$\begin{bmatrix} 0&0&0&0&0&0&0&2&2&4 \\ 0&0&0&0&0&0&0&0&1&2 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&1 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&2&1&3&0&0&0&0 \\ 0&0&0&1&1&2&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \end{bmatrix} \cong \begin{bmatrix} 0&0&0&0&0&0&0&1&1&1 \\ 0&0&0&0&0&0&0&0&0&1 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&1&1&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \end{bmatrix}$$

$$\overline{(A^2 \oplus A^4)A} \quad \cong \quad \overline{(A^2 \oplus A^4)A}$$

The resulting matrix, $\overline{(A^2 \oplus A^4)A}$, is the matrix of implied and observed correct and incorrect responses, which is then combined through Boolean addition with the original observed person-item and item-person responses, A.

$$\begin{bmatrix} 0&0&0&0&0&0&0&1&1&1 \\ 0&0&0&0&0&0&0&0&0&1 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&1&0&1&0&0&0&0 \\ 0&0&0&0&1&1&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \end{bmatrix} \oplus \begin{bmatrix} 0&0&0&0&0&0&1&1&0&1 \\ 0&0&0&0&0&0&0&1&1&1 \\ 0&0&0&0&0&0&0&1&0&1 \\ 0&0&0&0&0&0&0&0&1&1 \\ 0&0&0&0&0&0&0&0&0&1 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&1&1&1&1&1&0&0&0&0 \\ 0&0&0&1&0&1&0&0&0&0 \\ 0&0&0&0&1&1&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0&0 \end{bmatrix} = \begin{bmatrix} 0&0&0&0&0&0&1&1&1&1 \\ 0&0&0&0&0&0&0&1&1&1 \\ 0&0&0&0&0&0&0&1&0&1 \\ 0&0&0&0&0&0&0&0&1&1 \\ 0&0&0&0&0&0&0&0&0&1 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&1&1&1&1&1&0&0&0&0 \\ 0&0&0&1&1&1&0&0&0&0 \\ 0&0&0&0&1&1&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0&0 \end{bmatrix}$$

$$\overline{(A^2 \oplus A^4)A} \quad \oplus \quad A \quad = \quad (A^2 \oplus A^4)A \oplus A$$

Because it is possible to imply an answer to a person-item pair which contradicts the already observed response, a provision of the program at this stage prohibits such implications.

If we combine the binary matrix with person-person and item-item

relations with the new binary matrix of person-item and item-person relations we get an expanded persons + items x persons + items binary matrix which represents all our information. Each person and item has a row containing its dominances and a column with an entry for each person or item by which it is dominated. The row total minus the column total is a net dominance score which can be used to pair persons with items close to their level on the ability/difficulty continuum.

Ignoring guessing effects, an item at a person's level of ability has a .5 probability of being answered either correctly or incorrectly. This is the maximum degree of uncertainty that can exist about the outcome of a person-item confrontation. Items of greater or lesser difficulty and greater or lesser probability of being answered correctly are to some extent predictable, and, if we consider information to be the resolution of uncertainty, they represent lower information value. The richest source of information about person-item outcomes, then, is pairs of persons and items closely matched in ability/difficulty. Looked at in terms of the number of binary relations resolved, items which readily discriminate between adjacent persons in the person order create complete sets of relations for those persons. If two persons cannot be reliably ordered, then that binary relation will be missing from the overall matrix. For a given individual the most informative items are those which separate him from his closest neighbors. If we can differentiate each person from those closest to him, in the process we will collect the information necessary to differentiate him from everyone else.

We have already seen how dominance relations between persons and items, items and items, and persons and persons can be chained together to imply new relations, including implied correct and incorrect answers. The total number of binary relations within an ordered set of elements is $n(n - 1)/2$. In a binary dominance matrix containing $n \times n$ elements, each element x_{ij} has a complimentary element x_{ji} which expresses the same relation. The n diagonal elements are unable to express dominance, so we are left with $n(n - 1)/2$ elements. Most of these relations can be expressed in a variety of different ways. If we use the alphabet as an example, the matrix showing its binary order relations would have an entry for D follows A; entries for C follows A and D follows C; entries for B follows A and D follows B; or the set of B follows A, C follows B, and D follows C. Four sets of relations, then, tell us the order between A and D. Any two elements in the order can be implied from a number of chains of implications equal to the number of combinations possible using the intervening elements. Only the order of adjacent letters is not multiply determined. If the order of B and C is missing, there is no way to determine it from the remaining intersections. On the other hand, if we know the order of all $n - 1$ adjacent letter pairs the rest follow by implication. The order relations between adjacent or nearly adjacent elements can be used as building blocks to construct more distant relations, but the reverse is not the case. Distant relations provide little information about the other elements and are easily derived from multiple sources. For this reason items

and persons are chosen for their proximity on the ability/difficulty continuum.

On one hand we direct our observations of actual person-item relations to areas of greatest usefulness by matching the net dominance scores of persons and items. On the other hand distant relations in the person-item matrices are being filled in by implication. When the two processes converge and all person-item relations are determined, the test is over.

To summarize the implication process in matrix terms, consider the expanded matrix A, arbitrarily divided into submatrices:

$$A = \begin{bmatrix} I & \tilde{S}' \\ S & P \end{bmatrix}.$$

Where : I = item-item relations

P = person-person relations

S = person-item relations

$\tilde{S}'$ = transpose compliment of S (for complete data).

Observed correct and incorrect answers are recorded in S and $\tilde{S}'$ respectively:

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

I and P are null matrices.

Person-person and item-item relations are provided by AA.

$$AA = \begin{bmatrix} I & - \\ - & P \end{bmatrix} = \begin{bmatrix} (S\tilde{S}') & - \\ - & (\tilde{S}'S) \end{bmatrix}.$$

By significance testing I and P are transformed from integer products of S and $\tilde{S}'$ into binary matrices. Further person-person and item-item relations are implied from the squared matrices I and P. These also become binary after significance testing.

$$AAAA = \begin{bmatrix} I^2 & - \\ - & P^2 \end{bmatrix}$$

These implications are combined with the original item-item and person-person implications by Boolean addition.

$$AAAA \oplus AA = \begin{bmatrix} I \oplus I^2 & - \\ - & P \oplus P^2 \end{bmatrix}$$

The result is multiplied by the original entries in S and $\tilde{S}'$ and significance tested.

$$A(A^2 \oplus A^4) = \begin{bmatrix} - & \tilde{S}'(P \oplus P^2) \\ S(I \oplus I^2) & - \end{bmatrix}$$

The resulting binary implications are added, in Boolean addition, to the original person-item and item-person matrices with the provision that actual answers cannot be replaced or contradicted by implied relations.

$$A \oplus A(A^2 \oplus A^4) = \begin{bmatrix} - & \tilde{S}' \oplus \tilde{S}'(P \oplus P^2) \\ S \oplus S(I \oplus I^2) & - \end{bmatrix}$$

Although these matrix equations involve steps such as the reduction of integer matrices to binary matrices by significance testing and provisions for maintaining the original correct and incorrect responses, they can still be manipulated mathematically if the results are examined cautiously.

If we were dealing with simple matrix equations it can be noted that:

$$S(I \oplus I^2) = S(\tilde{S}'S \oplus \tilde{S}'S\tilde{S}'S) = S\tilde{S}'S \oplus S\tilde{S}'S\tilde{S}'S$$

$$\tilde{S}'(P \oplus P^2) = \tilde{S}'(S\tilde{S}' \oplus S\tilde{S}'S\tilde{S}') = \tilde{S}'S\tilde{S}' \oplus \tilde{S}'S\tilde{S}'S\tilde{S}'$$

Compare those results with two new equations:

$$(P \oplus P^2)S = (S\tilde{S}' \oplus S\tilde{S}'S\tilde{S}')S = S\tilde{S}'S \oplus S\tilde{S}'S\tilde{S}'S$$

$$(I \oplus I^2)\tilde{S}' = (\tilde{S}'S \oplus \tilde{S}'S\tilde{S}'S)\tilde{S}' = \tilde{S}'S\tilde{S}' \oplus \tilde{S}'S\tilde{S}'S\tilde{S}'$$

Therefore:

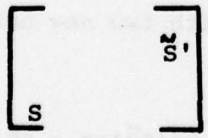
$$(P \oplus P^2)S = S(I \oplus I^2)$$

$$(I \oplus I^2)\tilde{S}' = \tilde{S}'(P \oplus P^2)$$

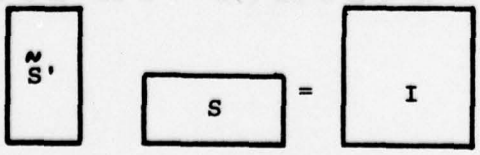
The information brought to the step of implying right and wrong answers by $(I \oplus I^2)$ and $(P \oplus P^2)$ would be equivalent except for the intervening processes just mentioned. There is a rough equivalence with the larger of the two matrices producing more implications when the program is operated with both matrix calculations.

For reasons of economy, only the item-item matrix is used in the implication process because it is generally larger than the person-person matrix, it maintains a constant size and can be reused with different persons. Person-person relations can still be derived after the person-item implications have been made by multiplying the final person-item matrix times the item-person matrix. The shortened procedure proceeds as follows:

Elements of S and $\tilde{S}'$ are observed:

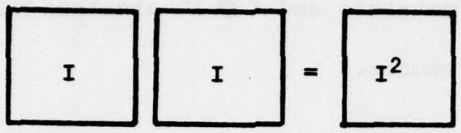


I is calculated: $\tilde{S}'S = I$



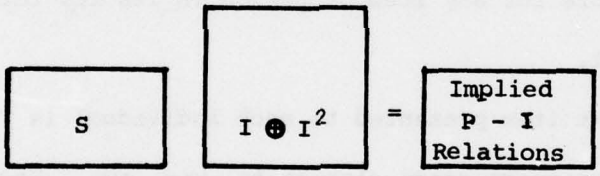
Each element i_{jk} is tested for significant dominance of a corresponding element i_{kj}. A binary entry is retained in I for each dominance.

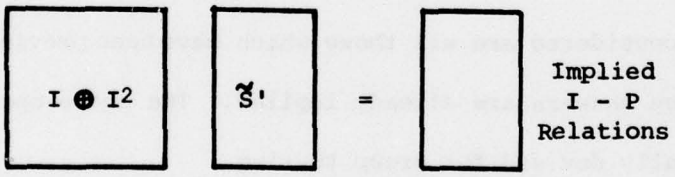
I is squared: $II = I^2$



The elements of I^2 are tested for significant dominances and the binary entries retained in I^2 are combined with those from I by Boolean addition.

The Boolean sum is then premultiplied by S to give implied person-item dominance relations and postmultiplied by $\tilde{S}'$ to give item-person dominance relations.

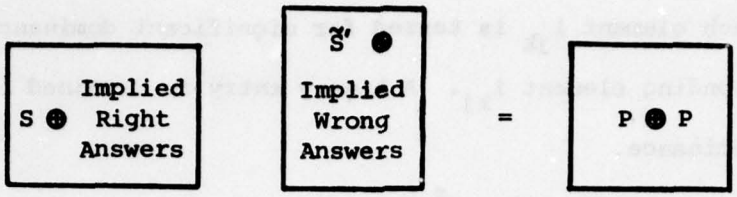




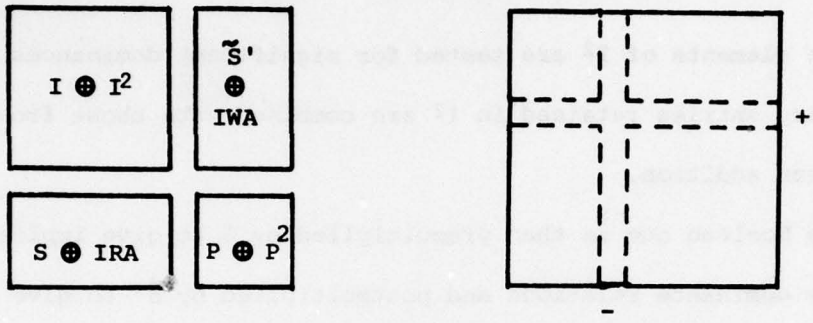
Each element in both implied dominance matrices is tested for significant dominance of its counterpart in the other matrix.

The binary results are added in Boolean fashion to S and $\tilde{S}'$ with the provision that actual answers are not contradicted by implied entries.

Person-person relations are then derived by significance testing the product of the implied and observed right and wrong answer matrices:



These three matrices and $I \oplus I^2$ are then used in the determination of new dominance scores.



The score for any item or person is its row total minus its column total.

The next item presented to each individual is the item which has a net dominance score closest to his own. Excluded from the items considered are all those which have been previously answered or whose answers are already implied. The above operations were originally devised for group testing.

This method of test tailoring has been under evaluation (Technical Report 4) using simulated group testing and a Fortran version of TAILOR (Cudeck, R. A., Cliff, N and Kehoe, J, 1977).

In a complex variety of circumstances, TAILOR produced an average correlation with true scores equal to .96 of the complete test correlation with true scores and used an average of 56% of the items.

TAILOR- APL

TAILOR-APL is not identical in its operation to the group testing

model. The differences consist mainly of computational shortcuts which are possible when information is being gathered for a single individual at a time. In a group test, information is gathered about item order quickly so that even the first person finished will take a test that has been tailored on the basis of other peoples' responses. The first person to take an individually administered test must answer all the items because there is no item information yet available. As will be shown later it may take only a few tests to begin extensive tailoring.

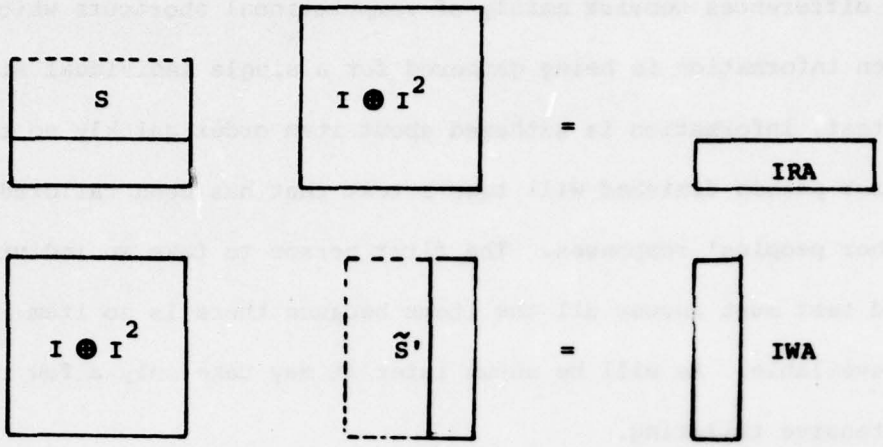
Individual testing is more appropriate than group testing for the largest potential application of tailored testing which is in conjunction with computer assisted instruction. McKillup and Urry (1976) of the U. S. Civil Service in their discussion of the advantages of computer administered tailored tests mention the ability to administer individual tests on a walk-in basis.

In implementing a version for individual testing the following economies seemed reasonable:

Because the impact of individual answers on the $I \oplus I^2$ matrix is likely to be small, it is only calculated at the end of each test before results are output. Also, during each test, implied answers, net dominance scores and implied person dominances based on P^2 are calculated only for the present examinee.

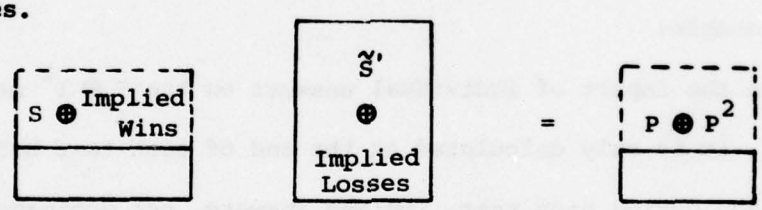
Directly observed correct and incorrect responses are recorded in S and $\tilde{S}'$.

$I \oplus I^2$ from the examinees already tested (a null matrix if this is the first test), is pre- and postmultiplied by the individual's vectors in S and $\tilde{S}'$ to obtain implied right and wrong answers after significance testing.

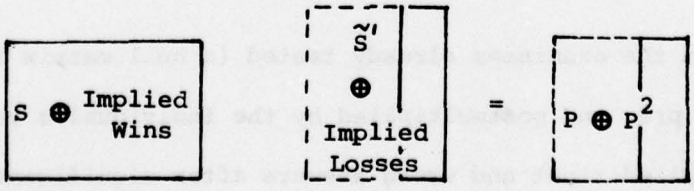


The significant binary counterparts of IRA (Implied Right Answers) and IWA (Implied Wrong Answers) are added to the individual's vectors in the versions of S and $\tilde{S}'$ that also contain implied responses, with the provision that actual answers are not contradicted.

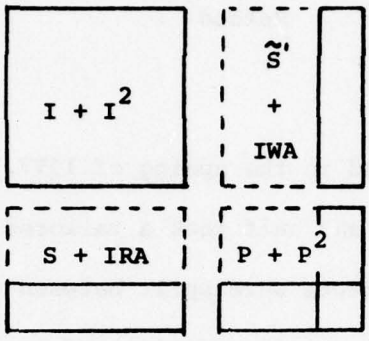
The individual's row of the S ● Implied Wins matrix is multiplied with the $\tilde{S}'$ ● Implied Losses matrix to give the individual's vector of integer person dominances.



The individual's column of the $\tilde{S}'$ ● Implied Losses matrix is multiplied with the S ● Implied Wins matrix to give the individual's vector of integers representing the number of times the person is dominated by other people.



After the row and column of $P \oplus P^2$ are tested against each other for significant dominance, the binary results are used in conjunction with the individual's vectors in S ⊕ Implied Wins and $\tilde{S}'$ ⊕ Implied Losses to derive the individual's net dominance score.



The individual's net dominance score is his row total minus his column total. Item scores from the last test are altered by the individual's entries in S ⊕ IRA and $\tilde{S}'$ ⊕ IWA. The only way the process differs from complete calculations is in not updating the entries of $I \oplus I^2$ until the test is finished.

TAILOR has been evaluated in the past with both Monte Carlo responses (Technical Report # 4) generated according to Birnbaum's model (Lord & Novick, 1968) and with test simulation using response matrices from previously administered complete tests. By generating responses from formulae it was possible to select levels of item discrimination, ability, difficulty, test length and other parameters with a precision and flexibility that real testing doesn't allow. Also, it is a good deal easier to arrange for a thousand simulated examinees than real ones. The reason for this experiment, the collection of data from real people, was to make sure the program which had been developed with artificial data would work as well with the real thing.

The current study was designed to compare tailored test reliability with the reliability of complete tests given under comparable conditions. The reliability of tailored scores combined with degree of test shortening would then demonstrate the measurement efficiency of TAILOR-APL.

Method

Design

Fifty subjects were tested in the spring of 1977. Half took a complete test administered by computer and half took a tailored test using the same item pool. Another fifty subjects were split between the experimental tailored test condition and the complete test condition and tested in the summer of 1977.

In all one hundred tests were given. Subjects were randomly assigned to the two conditions. Except for the number of items and the order in which they are presented, the tailored condition was identical to the complete test condition.

Due to the size constraints of the APL system at USC, the second group of tailored subjects did not take advantage of the stored information about item dominance provided by the first group. The second group, like the first began with no item information.

In order to obtain a measure of the reliability of tailored and non-tailored tests the item pool of 50 anagrams was divided randomly into two sets of items. These items were presented in an odd-even fashion; first an item from set one, then an item from set two. This picture was complicated slightly in the case of the tailored test. Because the length of the tailored tests cannot be predicted, the tailored halves were administered odd-even until one test was completed and then all remaining questions were from the

unfinished test.

Because the summer subjects didn't make use of item information gathered in the spring and because information about ability or difficulty is not shared between the split halves of the tailored test condition, there is a total of four tailored response matrices. In the presentation of results these four cases (and the corresponding cases for complete test data) are handled separately or merged, when appropriate, for the various analyses. The four matrices represent responses of the first and second 25 subjects to the A and B item pool halves.

Reliability was chosen as the principle criterion because it evaluates the tailored test and the complete test independently, unlike the criterion of tailored test correlation with complete test.

Correlation with complete test scores is appropriate only if we assume the items are independent and the answer an examinee gives to an item is not affected by the previous items presented. If that is true, a tailored test is simply a shorter and therefore less reliable version of the complete test. Using reliability allows for the possibility that tailored measures may better reflect the underlying proficiency being evaluated.

The individual testing version of TAILOR also allowed a second type of analysis to be performed. Because tailoring in the individual testing version is accomplished only to the extent that item information has accumulated from previous tests, the first tests include all the items, and subsequent tests show progressively greater influences from the tailoring procedure. The data therefore allows a regression analysis to be done using the order of administration as an independent variable which ranges from a complete test to the most tailored. If a significant trend toward higher or lower scores, more or less reliable scores or greater or less variance occurred this could be

detected by the regression analysis.

Subjects

Subjects were 100 students from the introductory psychology classes at the University of Southern California.

Items

Fifty unique solution anagrams were used in experimental and control conditions. The anagrams were taken from various sources and had either four or five letters (see Appendix A)[1]. Random arranging of the item order as well as the letter order was done by a separate APL program written by the investigator.

No information was gathered concerning item difficulty or discrimination before the experiment. The reason for using the anagrams was the ease of scoring answers by computer, rather than the existence of any statistical properties which would facilitate tailoring.

Procedure

Subjects were told the experiment was an evaluation of tailored testing and that they would be required to solve scrambled word problems presented by the computer at a typewriter style terminal. After answering any questions they had about the experiment and watching the first anagram appear, the experimenter left the room. Each anagram had a 30-second time limit. The time limit was the experimenter's estimate of a reasonable cutting point and was not based on any prior testing. When the test was finished a message was presented by the program telling the subject to notify the experimenter.

Scores

Three types of test scores are used. First a conventional number correct

was used for the complete test condition. Second a net dominance score was used internally for matching persons and items in the tailored tests and is also used in most of the analyses. Net dominance scores, as described earlier, involve subtracting the total number of elements (items and persons) which dominate a particular individual (or item) from the total of the elements which are dominated by that individual (or item). Third, in addition to net dominance scores, a score similar to the conventional number correct was computed for tailored test subjects to compare the tailored test score distributions to complete test scores. The difference between this second tailored test score and a simple correct answer score is that for a tailored test the score includes implied correct answers as well as .5 times the instances where an item is neither implied nor actually presented. Such missing entries are rare and no more than one ever occurred for a given individual.

Results

The obtained distributions of raw scores on both parallel forms of the anagrams test are displayed in Figures 3 and 4 for complete and tailored conditions. In addition to the net dominance tailored scores, scores calculated similarly to conventional scores, from the persons x items submatrix are also shown to allow a visual comparison of tailored and complete test distributions. The two sets of tailored scores are not just linear transformations, however, and the reliability of the net dominance scores is slightly higher.

The means obtained from raw tailored scores are 0.0800 and 2.3000 for the A and B parallel forms. Standard deviations for these distributions are 20.283 and 23.592. The two distributions are not significantly different by t-test ($\alpha = .6150$). The means of the recomputed tailored scores are 12.540 and 13.650. The corresponding means of the complete test forms are

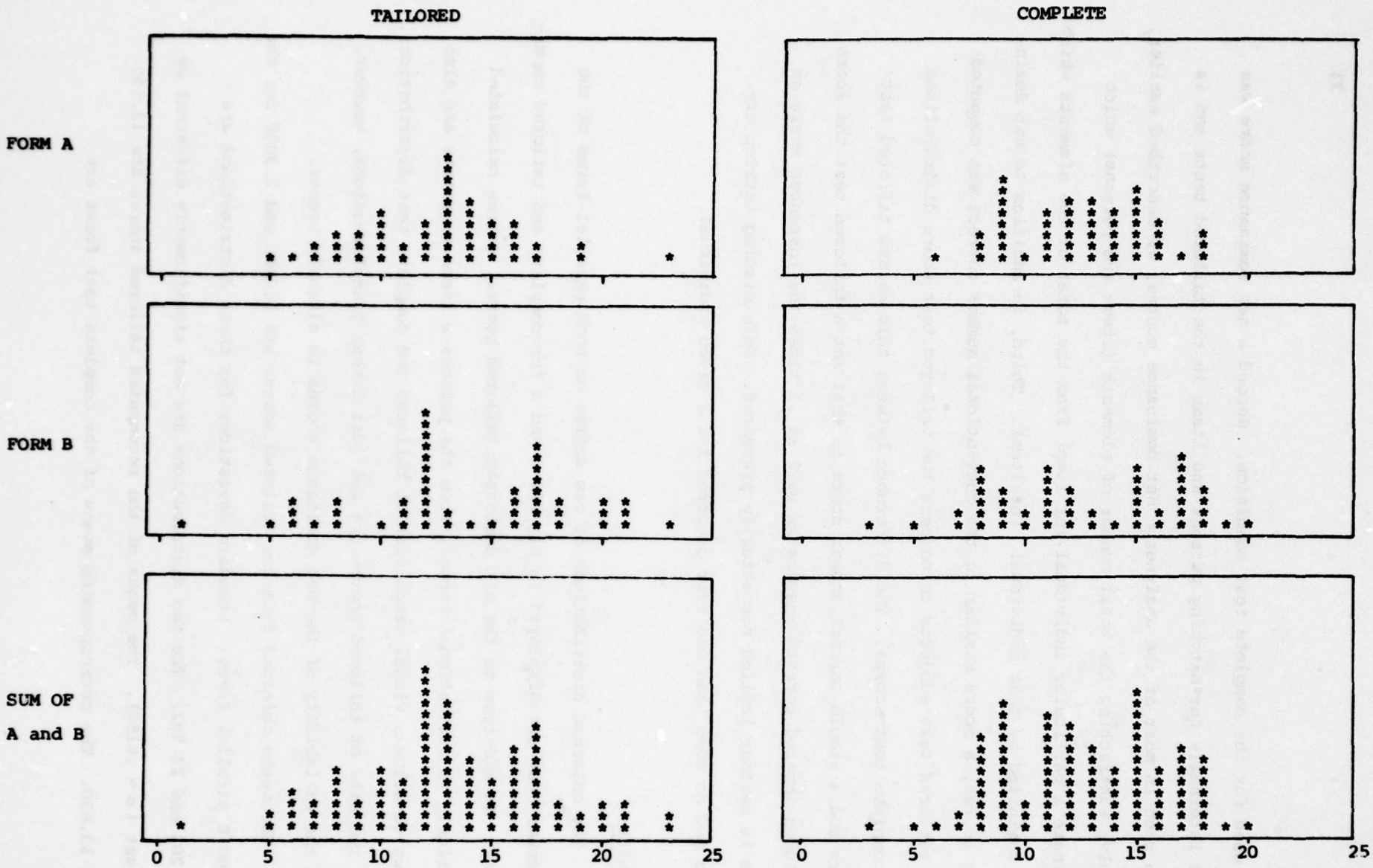


Figure 3: Histograms of complete test scores on Form A and Form B along with aggregated data and tailored scores calculated from a conventional persons by items matrix.

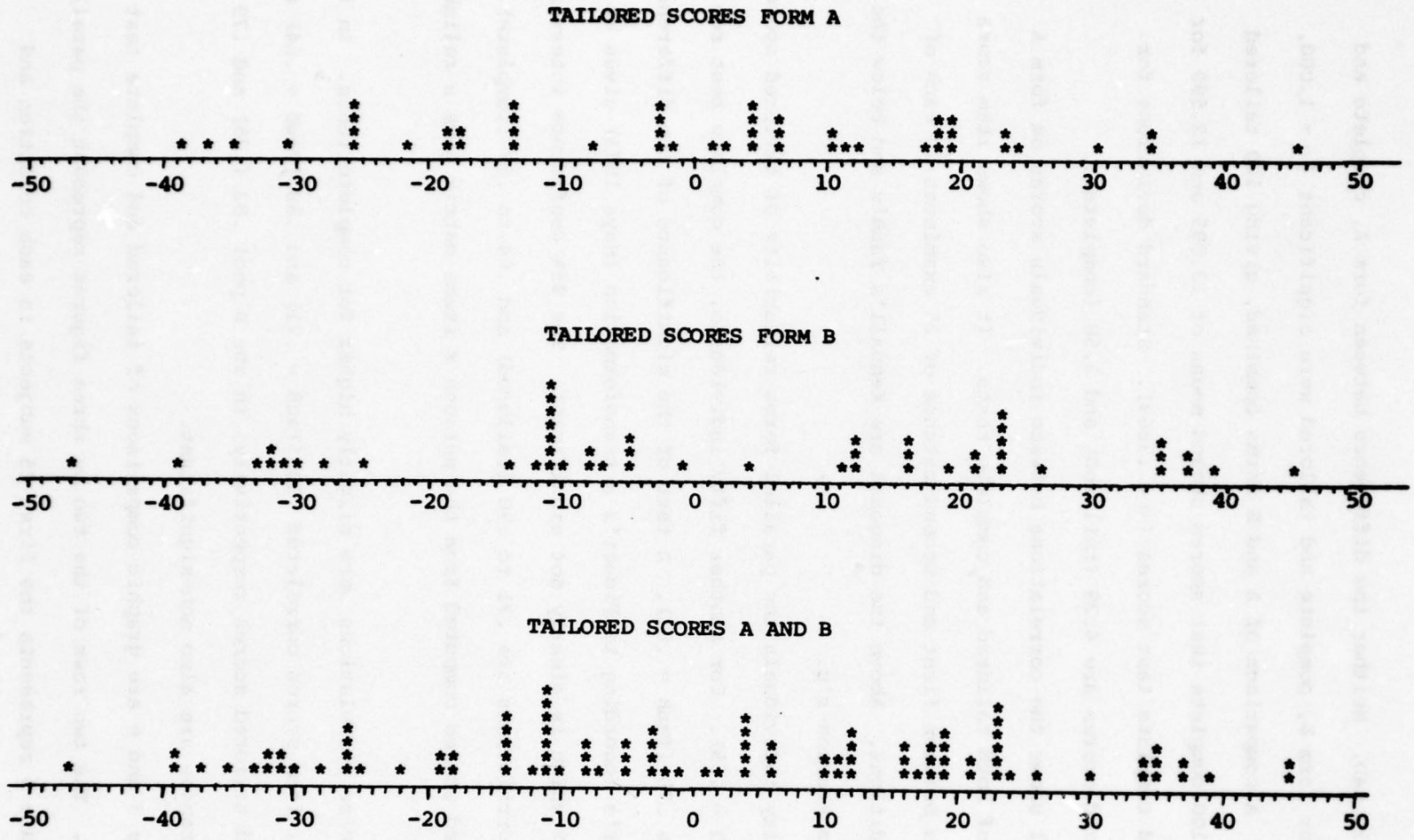


Figure 4: Histograms showing tailored score distributions, using net dominance scores.

12.540 and 12.640. Neither the difference between form A, complete and tailored, nor form B, complete and tailored were significant (α = 1.000, α = .2669). A comparison of A and B forms combined, giving 100 tailored scores and 100 complete test scores showed means of 13.095 and 12.590 for tailored and complete test scores (α = .3669). Standard deviations for these combined scores are 4.29 (tailored) and 3.58 (complete).

Table 1 shows the correlations between individuals scores on form A and form B of both tailored and complete tests. It also shows item score correlations between first and second batches of 25 examinees in each of the two conditions. Above the diagonal are Kendall's TauB's and below the diagonal are Pearson r's.

For fifty individuals the parallel forms reliability of tailored scores is .83 (TauB = .65). For another fifty individuals, the complete test reliability is .78 (TauB = .61). A test of the significance of the difference in Pearson r's according to Fisher's z transformation (Hays 1973) gives an alpha of .52 which is clearly not significant. The 95% confidence intervals for these correlations are .71 to .90 (tailored) and .64 to .87 (complete).

Tailored scores computed from the persons x items matrix gave a reliability of .79 .

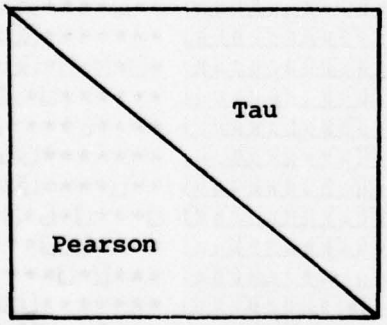
Item score correlations were slightly higher for complete tests. In the A item pool, item scores correlated .90 (TauB = .72) and .88 (TauB = .64) for complete and tailored scores respectively, in the B pool .83 (.75) and .79 (.58). These differences are also non-significant.

Figures 5 and 6 are graphic comparisons of tailored and complete test performance. The two rows of the two by three figures represent the parallel forms. Figure 5 represents the first 25 subjects in each condition and

Table 1

Item Correlations

Item Pools		A				B			
	Subject Groups	T1	T2	C1	C2	T1	T2	C1	C2
A	T1	--	.64	.72	.71	.67			
	T2	.88	--	.64	.73		.64		
	C1	.85	.92	--	.72			.51	
	C2	.91	.90	.90	--				.67
B	T1	.81				--	.58	.67	.60
	T2		.84			.79	--	.62	.69
	C1			.66		.80	.84	--	.75
	C2				.84	.87	.77	.85	--



Item Score Reliability	Person Score Reliability
Person Score Reliability	Item Score Reliability

T1 = first 25 tailored subjects
T2 = second 25 tailored subjects
C1 = first 25 complete test subjects
C2 = second 25 complete test subjects

Person Score Reliabilities for Combined data

Tailored r = .83 tauA = .65
Complete r = .78 tauA = .61

Tailored r based on scores from the conventional persons x items matrix

r = .79 tauB = .64

Figure 5: Response Matrices (First 25 Subjects)

Figure 6: Response Matrices (Second 25 Subjects)

Figure 6 the second 25. The first matrix in each row shows the right and wrong answers, represented by boxes and stars respectively, to questions which were actually presented to subjects. The second matrix in each row shows the actual responses and the responses which were implied. The blank spaces in column two represent implications which were revoked on the basis of information provided by later subjects, except for the final subject whose own responses revoked an implied answer when the $I \oplus I^2$ matrix was revised at the end of his test. The final matrix of each row is the right and wrong answers observed in each of the complete tests.

The rows and columns of each matrix are ordered by the person and item scores from that matrix only, so rows and columns are not comparable between matrices.

Figure 7 shows the observed correct and incorrect answers from the tailored tests arranged chronologically. Items are arranged according to difficulty. The top row of each matrix represents subject one's performance. Row two represents subject two and so on. The top two matrices are the A and B items and the first 25 subjects. The lower matrices represent the second 25 subjects.

From these illustrations we can see that the goal of clustering observations around an individual's ability level has been generally satisfied. It is also apparent that the longer vectors of observations were those gathered early in the experiment when the available information would not allow more extensive tailoring.

One of the most striking differences between tailored and complete tests is the greater uniformity of the second column of matrices in Figures 5 and 6 compared to the third column. The two regions of right and wrong answers are more cleanly separated for the tailored tests.

This lower degree of intermingling is a graphic display of two combined

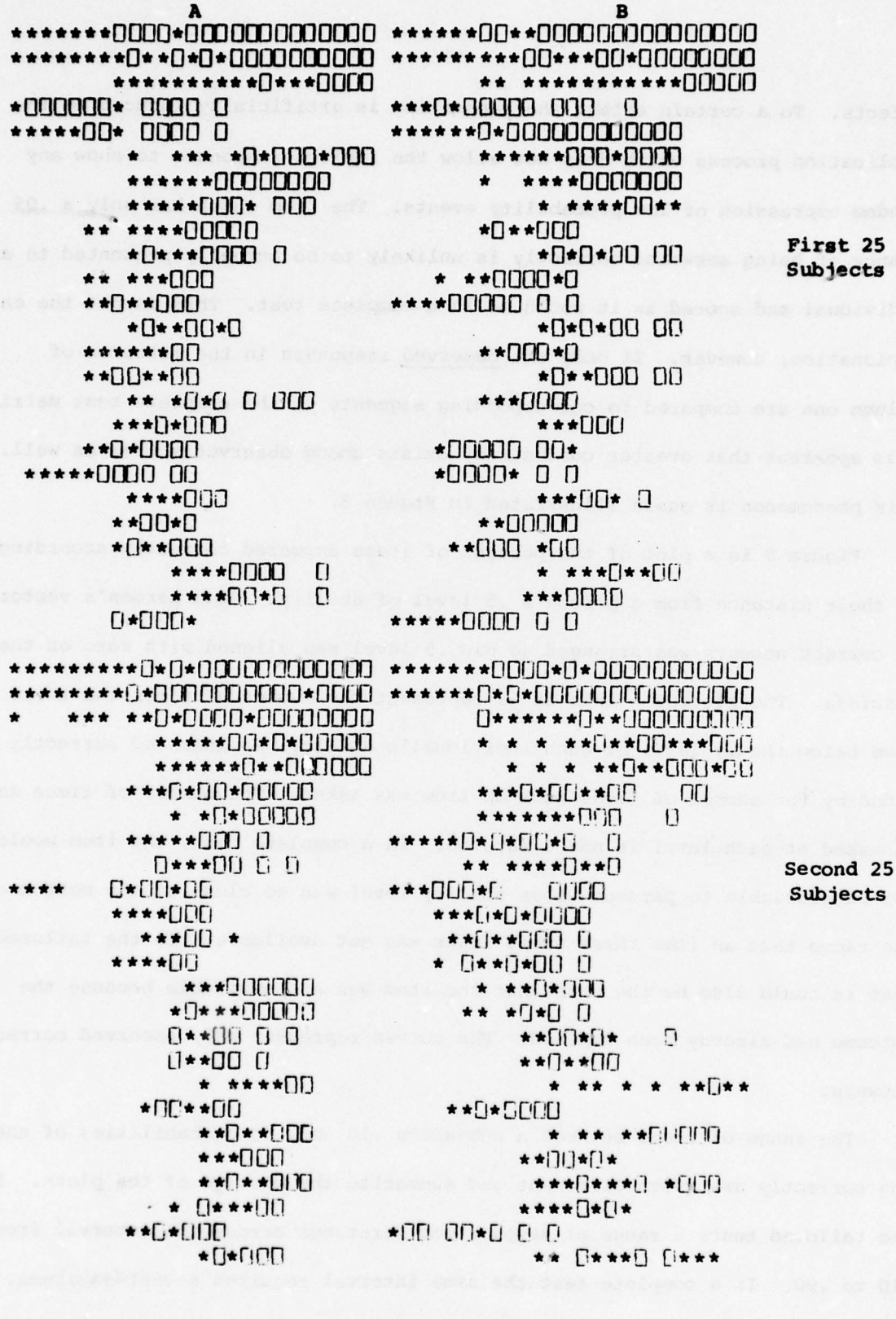


Figure 7: Observed Responses in Each Tailored Test Arranged Chronologically and According to Item Difficulty

effects. To a certain extent the separation is artificially induced by the implication process which does not allow the implied responses to show any random expression of low probability events. The item which has only a .05 chance of being answered correctly is unlikely to be actually presented to an individual and scored as it would be in a complete test. This is not the entire explanation, however. If only the observed responses in the matrices of column one are compared to corresponding segments of the complete test matrices, itis apparent that greater consistency exists among observed scores as well. This phenomenon is again illustrated in Figure 8.

Figure 8 is a plot of the percent of items answered correctly according to their distance from a person's .5 level of ability. Each person's vector of correct answers was arranged so his .5 level was aligned with zero on the abscissa. The plotted points at -3 represent the number of times the third item below the .5 level of each individual's ability was answered correctly divided by the number of times such an item was asked. The number of times an item is asked at each level is not always 50. In a complete test, the item would not be available to persons whose ability level was so close to the bottom of the range that an item three steps lower was not available. In the tailored test it could also be the case that the item was not available because the outcome had already been implied. The curves represent only observed correct answers.

The range of items between a subject's .10 and .90 probabilities of answering correctly are quite different and summarize the message of the plots. In the tailored tests a range of seven items stretches across the interval from .10 to .90. In a complete test the same interval requires seventeen items. This difference is also expressed by the consistency indices computed for the

tailored and complete test, observed response matrices. The average value of C_{t3} (Cliff, 1977) for tailored tests is .42. The average for the complete tests is .16.

Figure 9 shows the number of questions asked, plotted as a function of the number of tests given. An average of forty-four percent of the questions were presented in each tailored test. The range and mean are presented for each position in the order.

The curve in Figure 9 appears to asymptote near eight items after fifteen tests have been administered. The average of the last ten tests in all four tailored conditions is 7.975 items asked. If we continued to give the anagrams test we could expect to present an average of 32 percent of the items.

As previously mentioned, the progression of the tailoring process from a complete test to about eight items allows us to use serial position as an independent regression variable representing the extent of tailoring.

Regressions were done with z scores, absolute values of z scores and the difference of z scores on forms A and B as the dependent variables. It was intended that a trend in z scores would show if tailoring induced higher or lower test scores. The correlation obtained was a non-significant -.067 so details of the regression are not presented. The regression of absolute z scores was intended to detect any change in variance that might occur because of tailoring. This correlation was 0.000. The differences between z scores were analyzed to show any tendency for tailoring to change the reliability of the test. The correlation again was non-significant ($r = .114$). Thus scores from later subjects based on 8 items seem to be as reliable as scores from earlier subjects who were given much longer tests.

Figure 10 shows the actual branching of two examinees within the A and B item pools. The items were ordered from easiest to hardest, left to right.

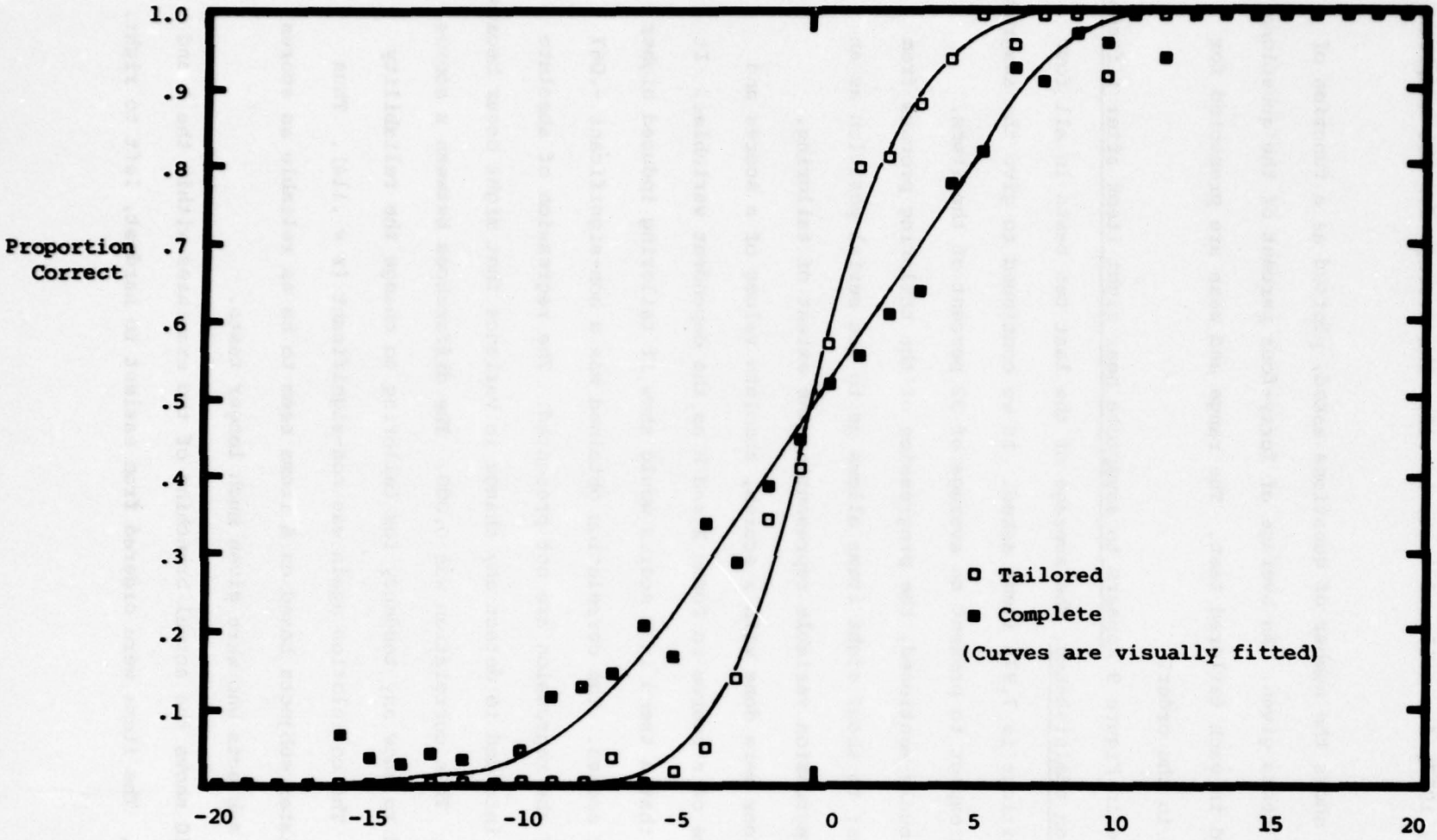


Figure 8: Proportion of correct responses in Tailored and Complete tests for items of a particular distance from the individual's estimated ability.

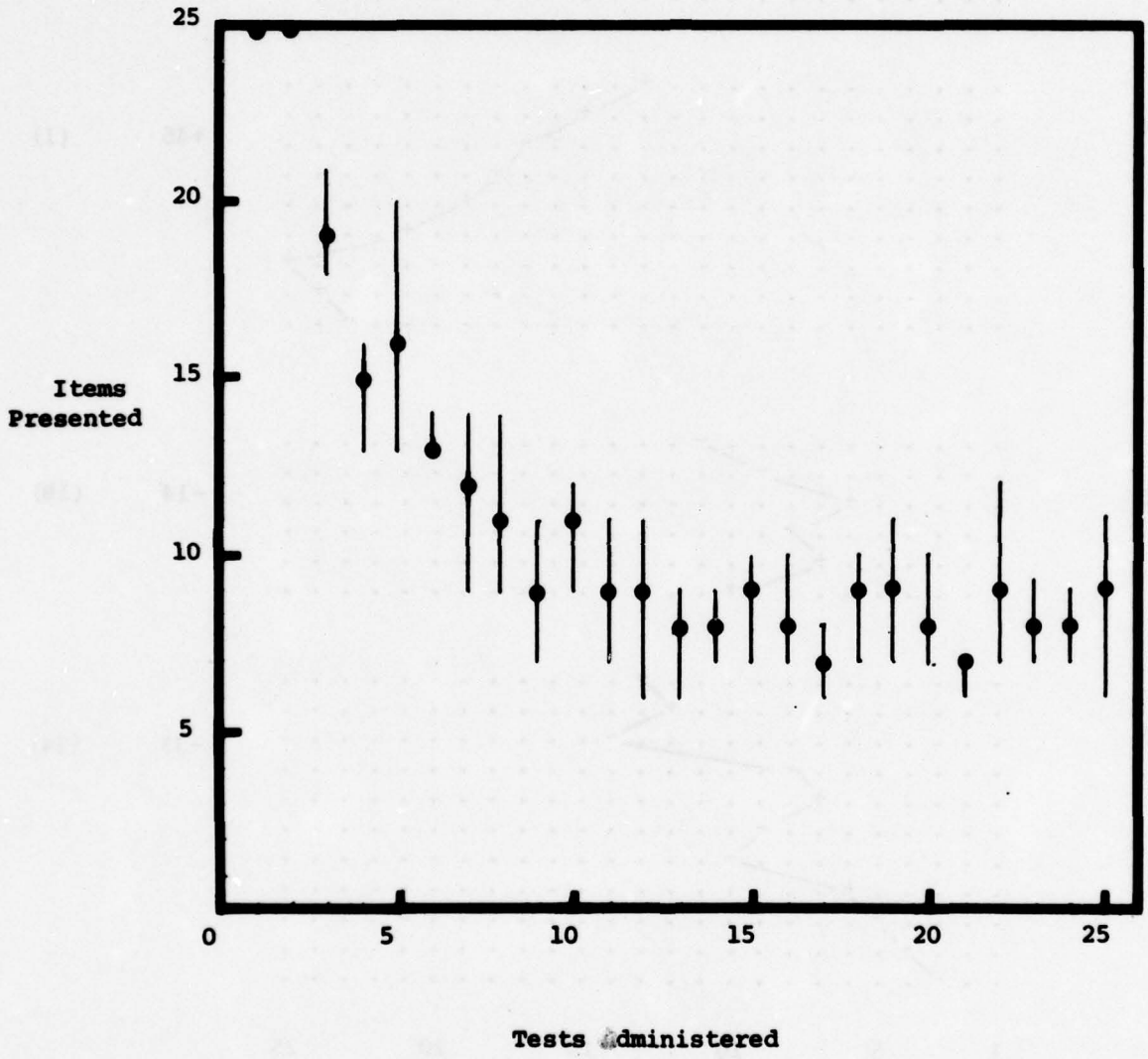


Figure 9: Range and mean number of items presented during four applications of TAILOR-APL plotted by the number of persons tested in each block of 25 subjects.

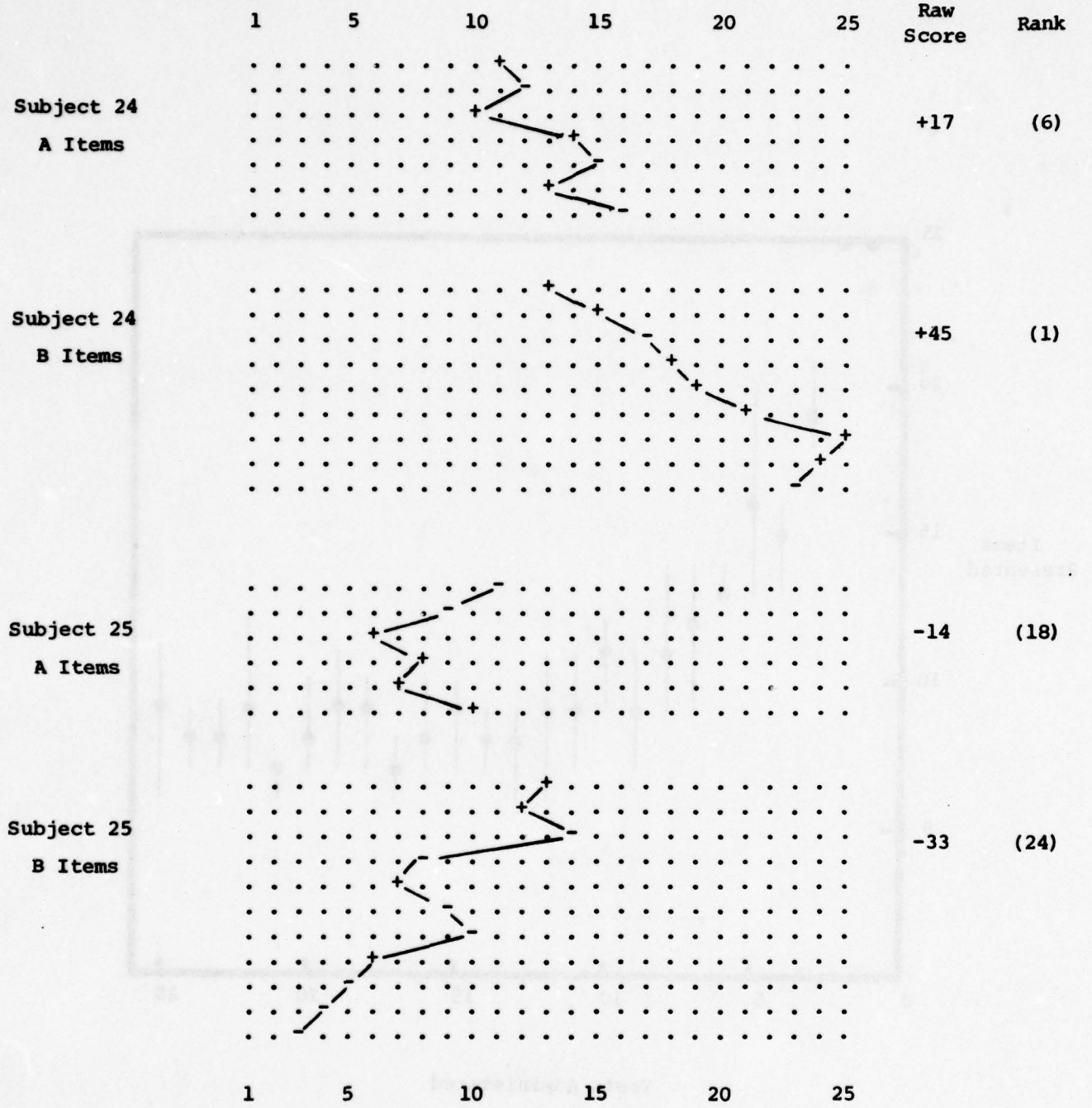


Figure 10: Examples of two individuals being routed through the A and B item pools by TAILOR-APL

score is closest to zero. Plusses and minuses in the chain represent correct and incorrect answers as the test progresses. After each response, both items and persons receive new scores and the examinees are assigned to the items which best match their net dominance scores and which haven't been given or implied yet. At the end of each test all the items have either implied answers or have actually been given and scored.

From the figure it is possible to see that the size and direction of the branching step is variable. Wrong answers are not necessarily followed by easier items or correct answers by harder items although that is, in general, the case.

Although this pair of examinees took many more items and received much more extreme scores in the B item pool than in the A item pool, this is not generally true to such an extent. The B items over all subjects involved only 3% more items in an average test and the standard deviations of B and A scores were 23.592 and 20.283 respectively.

Discussion

Based on previous evaluations of TAILOR in the group testing form with Monte Carlo techniques and simulation testing, there was reason to expect the individual testing version to produce reliabilities slightly less than complete tests and reduce the items presented to a half or a third depending on the quality of the item pool.

The overall level of tailoring (.44 of the items presented) and the asymptotic value (.32) were therefore, not unexpected, but the speed with which tailoring took place in terms of examinees was very surprising. In each of four tailored tests only two examinees had to answer 25 items. After 15

subjects had been tested the program seemed to reach an asymptotic value. This contrasts rather startlingly with tailoring techniques that require 100-150 complete tests (in the case of Reckase's one parameter method), up to several thousand recommended frequently for the more complex procedures (summary in Reckase 1977).

Had TAILOR-APL produced scores not significantly less reliable than complete test scores, the program would have done what could reasonably be expected of a test tailoring method, but again TAILOR-APL outperformed our expectations. Though Fisher's z transformation fails to show a significant increase in reliability, the overall reliability is necessarily a compromise between the shortening of the test and what appears to be an increase in the discrimination of the items. If the remaining items had been added to the end of the tailored test the difference in reliabilities may have been significant.

Although the data in Figure 8 seem to be dramatically different for items in the tailored and complete test condition, the writer confesses his inability to test the difference in slope for significance.

For some reason, the same anagrams when presented in a tailored test are delineating more precisely between ability levels. There are reasons why a tailored test might provide more reliable measurement than a conventional test. By presenting items close to a person's ability, guessing is not encouraged as it would be if the items were too difficult, nor are examinees likely to become bored and careless as they might if the test was felt to be too easy. It is possible that by requiring a steady effort from the examinee his behavior is made more consistent.

An additional influence may be TAILOR's tendency to begin each test with

items either too difficult or too easy and slowly drift past the examinee's ability level. Consistency could be induced by the elevating or depressing effects of early items on later performance.

If we assume that as long as an examinee is performing well his responses to test items are better than they would normally be, he will tend to pass his usual level of performance before making an error. When an error is finally made the elevating effect is ended and items beyond that point are consistently failed as they would have been normally. Since not all items presented after the initial failure will be above the normal ability level, the discrimination will not be perfect, but the tendency of this process will be to increase discrimination. A similar effect could be hypothesized in the opposite direction due to a depressing effect caused by too difficult items. To distinguish between these hypotheses a tailoring system which presented items near a person's ability level, but not in order of difficulty, would have to be compared to the present system.

There are three issues concerning the operation of TAILOR-APL which are beyond the current investigation and which may be important in the future. The first two are potential difficulties which were not troublesome in the current study, but which may cause problems in a new application. The third is a potential difficulty in <u>other</u> tailoring programs which will illustrate some of the unique benefits of basing the tailoring process on information gathered during the administration of tailored tests.

There is a possibility that if the items are not as evenly divided by the mean ability of the examinees, as they are here, giving a new examinee the item whose net dominance score is closest to zero will not be the item closest to a .50 probability of being answered correctly. Because the item

score includes item dominance information as well as person dominances, an item which has a net person dominance of zero will be offset from zero according to its item dominances.

This problem is peculiar to the individual testing version of TAILOR, TAILOR-APL. The original formulation of the technique (Cliff, 1975) was for the group testing version in which both items and persons collect dominances simultaneously so items will not be offset by the results of previously administered tests. Any item dominances will be equally attached to person and item scores as the test progresses.

Another potential problem in the operation of the program concerns the treatment of poorly discriminating items. The program is designed to continue a test until all item outcomes are implied or observed through actual administration. A poorly discriminating item is the least likely item to be implied since it has the fewest reliable dominance relations with other items. The result could be that the worst items in the item pool are asked most often simply becuase they have few of the relations to the rest of the test which would allow them to be implied.

The mechanism for eliminating poorly discriminating items has not yet been implemented, but the derivation of indices which will be necessary for this task has been accomplished (Cliff, 1977).

A grooming of the item pool to eliminate ability/difficulty offset could also be included in the subroutine which removes non-discriminating items. Once indices are computed and the mean ability level is known with reasonable accuracy both tasks could be accomplished by little more than visual inspection or a simple cutoff value.

The advantages mentioned above of using tailored test item information as a basis for tailoring testing are based on two types of error that are intro-

duced in parameters estimated from complete test data.

First there are errors introduced because of order effects among the items. Since the order in which items are presented in tailored tests differs systematically from the order in which they are presented in a complete test, any change in item discrimination or difficulty caused by the order in which items are presented during tailored testing will result in inaccuracy and inefficiency in tailoring systems which rely on accurate parameter estimates.

The greater discrimination of the items in the current study when presented in a tailored test is evidence that such order effects do occur. If the higher discrimination is a general characteristic of tailored tests there will be other studies which will find higher reliabilities in tailored tests than in complete tests. Tailored test reliability can only be higher than complete test reliability if the items are not independent of each other. If the examinee's history of correct and incorrect answers is effecting the probability of his correctly answering the current item, then the model which says the probability of a correct response to a particular item is purely a function of that examinee's true score is not accurately describing the examinee's behavior. Because the current data is not a statistically significant example of higher reliability it must by itself remain only suggestive.

Killcross (1976) in his review of tailored testing recognized the importance of item order effects and suggested an index of context reliability be devised. He reviews nine studies which looked at such things a whether students would score higher on a test which was given in ascending item difficulty or descending difficulty, whether overall test variance and mean difficulty were accurately reflected by data from small item samples and whether success of failure on an item caused the next item to be harder or easier.

None of the studies, apparently, used speeded items, items with knowledge of results provided or subsets of items ordered in difficulty or matched to the examinee's ability. The only positive finding was that items in a 'quantitative thinking' test became significantly easier when the test was reduced to one fourth its size.

Although some elements of the way items were presented in tailored testing were looked at in these studies, they were never looked at all together or even all separately, nor was discrimination used as a criterion. There are indications from tailored testing research done elsewhere that order effects are operating.

Waters (1976) reported higher validities for "stradaptive" tests with 19, 25 or 31 vocabulary items compared to a conventional test of 48 items. The number of items in a "stradaptive" test is to some extent deceptive in this case because an initial graded response item is used to determine the individual's entry point. This may have an effect equivalent to several items administered within the test itself. The graded response item is unlikely to have an effect equal to the 10 or 20 items which represent the differences between the tailored and the complete tests.

Waters obtained correlations with a criterion of .499, .536, and .536 for the tailored tests. The correlation between the complete test and the criterion was .477.

Not all tailoring techniques would be adversely effected by such alterations in discrimination parameters, if they become a familiar result of tailoring. Certainly they had no harmful effects on the "stradaptive" procedure. Alteration of other item characteristics, however, seem even more likely. A series of investigations at Minnesota (Betz & Weiss, 1976a and 1976b; Weiss

1975) shows that when knowledge of results is provided during tailored testing dramatic changes occur in item difficulty. Since it is impossible for knowledge about the correctness of one's answer to affect performance on the item for which it is given, this necessarily means the items are not independent.

Aside from the problem of item parameters changing when the items are moved from a conventional test to a tailored test, there is the additional problem of the appropriateness of the parameters and standards for any particular sample. It may be argued that changing a test for every group that takes it destroys the soundness of the measurements as standardized evaluations. In fact, however, careful re-evaluation of items is necessary for the same reasons that the intial item analysis takes place during construction of the test. A score on the Stanford-Binet taken in 1960 can be as much as ten IQ points away from the same score achieved by a person of identical age in 1972 (Sattler, 1974). A vocabulary item which discriminates well when given in an intelligence test in one population may give poor discrimination when administered to people in another.

The sheer burden of collecting hundreds or thousands of complete test responses to re-parameterize a tailored test makes the tailoring methods which gather their information from complete test data more prone to being administered with parameters that are out of date or inappropriate, assuming, of course, that appropriate parameters can be derived from complete test data to begin with.

Pretesting is not only an expensive undertaking (Reckase, 1977) which would make tailored testing infeasible for tests which will not be given to thousands of examinees, but it also jeopardizes item security and requires a double standard in testing. Some examinees get a tailored test and others

have to spend time with a longer examination and face the possiblity of a less reliable score.

In contrast to the pretesting tailoring methods, TAILOR quickly adapts to any changes in person-item relations. It can also be "frozen" with a particular reference sample is the user wishes. Item security is not jeopardized by using the same item pool year after year nor is the operation of the procedure jeopardized by "knowledge of results" if the tester chooses to give his students feedback.

Program Availability

Copies of the APL program used for this research are available from Douglas McCormick, Psychology Department, University of Southern California, University Park, Los Angeles, California, 90007.

Conclusion

This first empirical test of TAILOR-APL with 50 live examinees has shown that a test can be reduced to 44% of its original length with no pretesting of the items and no significant loss in test reliability when comparison is made to a complete test administered under comparable conditions.

Reference Notes

Angoff, W. H. and Huddleston, E. M. The multi-level experiment: A study of a two-level test system for the College Board Scholastic Aptitude Test. (Statistical Report SR-58-21). Princeton, New Jersey: Educational Testing Service, 1958.

Betz, N. E. and Weiss, D. J. Effects of Immediate Knowledge of Results and Adaptive Testing on Ability Test Performance. (Technical Report 76-3). Minneapolis, Minnesota: University of Minnesota, Dept of Psychology, 1976. (a)

Betz, N. E. and Weiss, D. J. Psychological Effects of Immediate Knowledge of Results and Adaptive Ability Testing. (Technical Report 76-4). Minneapolis, Minnesota: University of Minnesota, Dept of Psychology, 1976. (b)

Cliff, N., Cudeck, R. A. and McCormick, D. J. Evaluations of Implied Orders as a Basis for Tailored Testing Using Simulations. (Technical Report 4). Los Angeles, California: University of Southern California, Dept of Psychology, 1977.

Ferguson, R. L. Computer-assisted criterion-referenced testing. (Working Paper No. 49). Learning Research and Development Center, University of Pittsburgh, 1969.

Ferguson, R. L. and Hsu, T. The application of item generators for individualizing measurement. (Report No. 14). Learning Research and Development Center, University of Pittsburgh, 1971.

Killcross, M. C. A review of research in tailored testing, Report APRE No. 9/76. Ministry of Defense, Army Personnel Research Establishment, Royal Aircraft Establishment, Farnborough, Hants., Great Britain, 1976.

Mussio, J. J. A Modification of Lord's model for tailored tests. Unpublished doctoral dissertation, University of Toronto, 1972.

Novick, M. R. Bayesian methods in psychological testing. (Research Bulletin RB-69-31). Princeton, New Jersey: Educational Testing Service, 1969.

Owen, R. J. A Bayesian approach to tailored testing. (Research Bulletin RB-69-92). Princeton, New Jersey: Educational Testing Service, 1969.

Owen, R. J. Bayesian sequential design and analysis of dichotomous experiments with special reference to mental testing. Unpublished manuscript. Dept of Statistics, University of Michigan, 1970.

Reckase, M. D. Ability estimation and item calibration using the one and three parameter logistic models: a comparative study, Research Report 77-1, Tailored Testing Research Laboratory, Educational Psychology Dept, University of Missouri, Columbia, MO., November 1977.

Urry, V. W. A Monte Carlo investigation of logistic test models. Unpublished doctoral dissertation, Purdue University, 1970.

Weiss, D. J. Computerized Adaptive Trait Measurement: Problems and Prospects. (Technical Report 75-5). Minneaplis, Minnesota: University of Minnesota, Dept of Psychology, 1975.

Weiss, D. J. and Betz, N. E. Ability Measurement: Conventional or adaptive? (Technical Report 73-1). Minneapolis, Minnesota: University of Minnesota Dept of Psychology, 1973.

Wood, R. The application of Bayesian sequential analysis to educational and psychological testing. Paper presented at the meeting of the American Educational Research Association, Minneapolis, Minnesota, March 1970.

References

Bradway, K. P. Comparison of standard and wide-range testing on the Stanford-Binet. Journal of Consulting Psychology, 1943, 7, 179-182.

Cleary, T. A., Linn, R. L., and Rock, D. A. An exploratory study of programmed tests. Educational and Psychological Measurement, 1968, 28, 345-360.(a)

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

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

Cliff, N. A theory of Consistency of Ordering Generalizable to Tailored Testing. Psychometrika, 1977, 42, 375-399.

Cowden, D. J. An application of sequential sampling to testing students. Journal of the American Statistical Association, 1946, 41, 547-556.

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

Dunn, L. M. Expanded manual for the Peabody Picture Vocabulary Test. Minneapolis: American Guidance Service, 1965.

Gugel, J. F., Schmidt, F. L., and Urry, V. W. Effectiveness of the ancillary estimation procedure. Proceedings of the First Conference on Computerized Adaptive Testing, June 1975, 1976, 103-106.

Guilford, J. P. and Fruchter, B. Fundamental statistics in Psychology and Education. New York: McGraw-Hill, 1973.

Hansen, D. H. Reflections on adaptive testing. Proceedings of the First Conference on Computerized Adaptive Testing, June 1975, 1976, 90-94.

Hays, W. L. Statistics for the Social Sciences. New York: Holt, Rinehart and Winston, 1973.

Hick, W. E. Information theory and intelligence tests. British Journal of Psychology, Statistical Section, 1951, 4, 157-164.

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

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

Lord, F. M. The Self-scoring Flexilevel Test. Journal of Educational Measurement, 1971, 8, 147-151. (b)

Lord, F. M. and Novick, M. R. Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley Publishing Company, Inc. 1968.

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

McKillup, R. H. and Urry, V. W. Computer-assisted testing: An orderly transition from theory to practice. Proceedings of the First Conference on Computerized Adaptive Testing, U. S. Civil Service Commission, Washington, D. C., 1976.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute of Educational Research, 1960.

Reckase, M. D. An interactive computer program for tailored testing based on the one-parameter logistic model. Behavior Research and Instrumentation, 1974, 6, 208-212.

Sattler, J. M. Assessment of Children's Intelligence. Philadelphia, London, Toronto: W. B. Sanders Company, 1974.

Wald, A. Sequential Analysis. New York, New York: Wiley, 1947.

Waters, B. W. An empirical investigation of Weiss' Stradaptive Testing Model. Proceedings of the First Conference on Computerized Adaptive Testing, June 1975, 1976, 54-63.

Wechsler, D. The measurement and appraisal of adult intelligence (4th ed.). Baltimore: Williams and Wilkins, 1958.

Footnote

1. Many came from Gilhooly, K. J. and Hay, D. Imagery, concreteness, age-of acquisition, familiarity and meaningfulness values for 205 five-letter words having single solution anagrams. Behavioral Research Methods, 9(1), 1977.

Appendix A: Anagrams and Solutions

Set A		Set B	
sifh	fish	hymer	rhyme
ihpmc	chimp	naldg	gland
dtovi	divot	owrk	work
ofctr	croft	abclk	black
ilyrc	lyric	tnoek	token
ucont	count	ilpa	pail
firya	fairy	culnh	lunch
gryol	glory	uavlt	vault
pleh	help	odnf	fond
ihbrc	birch	ecnbh	bench
paonr	apron	nevah	haven
hglit	light	albez	blaze
letl	tell	knale	angle
opitv	pivot	ihra	hair
ibta	bait	nogme	gnome
kool	look	ibrot	orbit
orfo	roof	lsed	sled
htpde	depth	rlveo	lover
fkeni	knife	carhi	chair
enuco	ounce	goibt	bigot
oolw	wool	epil	pile
ubdto	doubt	evcor	cover
mdone	demon	krnac	crank
tlave	valet	bdran	brand
omnaw	woman	enog	gone

Distribution List

Navy

4 Dr. Marshall J. Farr, Director
Personnel and Training Research Programs
Office of Naval Research (Code 458)
Arlington, VA 22217

1 ONR Branch Office
495 Summer Street
Boston, MA 02210
ATTN: Dr. James Lester

1 ONR Branch
1030 East Green Street
Pasadena, CA 91101
ATTN: Dr. Eugene Gloye

1 ONR Branch Office
536 South Clark Street
Chicago, IL 60605
ATTN: Dr. Charles E. Davis

1 Dr. M. A. Bertin
Scientific Director
Office of Naval Research
Scientific Liaison Group/Tokyo
American Embassy
APO San Francisco 96503

1 Office of Naval Research
Code 200
Arlington, VA 22217

6 Director
Naval Research Laboratory
Code 2627
Washington, DC 20390

1 Technical Director
Navy Personnel Research and Development Center
San Diego, CA 92152

1 Assistant for Research Liaison
Bureau of Naval Personnel (Pers Or)
Room 1416, Arlington Annex
Washington, DC 20370

1 Assistant Deputy Chief of Naval Personnel for Retension Analysis and Coordination (Pers 12)
Room 2403, Arlington Annex
Washington, DC 20370

1 CDR Paul D. Nelson, MSC, USN
Naval Medical R & D Command (Code 44)
National Naval Medical Center
Bethesda, MD 20014

1 Commanding Officer
Naval Health Research Center
San Diego, CA 92152
ATTN: Library

1 Chairman
Behavioral Science Department
Naval Command & Management Division
U. S. Naval Academy
Annapolis, MD 21402

1 Dr. Jack R. Borsting
U. S. Naval Postgraduate School
Department of Operations Research
Monterey, CA 93940

1 Director, Navy Occupational Task Analysis Program (NOTAP)
Navy Personnel Program Support Activity
Building 1304, Bolling AFB
Washington, DC 20336

1 Office of Civilian Manpower Management
Code 64
Washington, DC 20390
ATTN: Dr. Richard J. Niehaus

1 Superintendent
Naval Postgraduate School
Monterey, CA 93940
ATTN: Library (Code 2124)

1 Chief of Naval Technical Training
Naval Air Station Memphis (75)
Millington, TN 38054
ATTN: Dr. Norman J. Kerr

1 Principal Civilian Advisor
for Education and Training
Naval Training Command, Code 00A
Pensacola, FL 32508
ATTN: Dr. William L. Maloy

1 Director
Training Analysis & Evaluation Group
Code N-00t
Department of the Navy
Orlando, FL 32813
ATTN: Dr. Alfred F. Smode

1 Navy Personnel Research
and Development Center
Code 01
San Diego, CA 92152

5 Navy Personnel Research
and Development Center
Code 02
San Diego, CA 92152
ATTN: A.A. Sjoholm

2 Navy Personnel Research
and Development Center
Code 310
San Diego, CA 92152
ATTN: Dr. Martin F. Wiskoff

1 Navy Personnel Research
and Development Center
San Diego, CA 92152
ATTN: Library

1 Navy Personnel Research
and Development Center
Code 9041
San Diego, CA 92152
ATTN: Dr. J. D. Fletcher

1 D. M. Gragg, CAPT, MC, USN
Head, Educational Programs Development Department
Naval Health Sciences Education and Training Command
Bethesda, MD 20014

Army

1 Technical Director
U. S. Army Research Institute for the Behavioral and Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

1 Headquarters
U. S. Army Administration Center
Personnel Administration Combat Development Activity
ATCP- HRQ
Ft. Benjamin Harrison, IN 46249

1 Armed Forces Staff College
Norfolk, VA 23511
ATTN: Library

1 Dr. Stanley L. Cohen
U. S. Army Research Institute for the Behavioral and Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

1 Dr. Ralph Dusek
U. S. Army Research Institute for the Behavioral and Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

1 Dr. Joseph Ward
U. S. Army Research Institute for the Behavioral and Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

1 HQ USAREUR & 7th Army
ODCSOPS
USAREUR Director of GED
APO New York 09403

1 ARI Field Unit - Leavenworth
Post Office Box 3122
Fort Leavenworth, KS 66027

1 Dr. Ralph Canter
U. S. Army Research Institute for the Behavioral and Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

1 Dr. Milton Maier
U. S. Army Research Institute
for the Behavioral and Social
Sciences
1300 Wilson Boulevard
Arlington, VA 22209

1 Dr. Milton S. Katz, Chief
Individual Training & Perfor-
mance Evaluation
U. S. Army Research Institute for
the Behavioral and Social
Sciences
1300 Wilson Boulevard
Arlington, VA 22209

Air Force

1 Research Branch
AF/DPMYAR
Randoph AFB, Tx 78148

1 Dr. G. A. Echstrand (AFHRL/AST)
Wright Patterson AFB
Ohio 45433

1 AFHRL/DOJN
Stop #63
Lackland AFB, TX 78236

1 Dr. Martin Rockway (AFHRL/TT)
Lowry AFB
Colorado 80230

1 Dr. Alfred R. Fregly
AFOSR/NL
1400 Wilson Boulevard
Arlington, VA 22209

1 AFHRL/PED
Stop #63
Lackland AFB, TX 78236

1 Major Wayne S. Sellman
Chief of Personnel Testing
HQ USAF/DPMYP
Randolph AFB, Tx 78148

Marine Corps

1 Director, Office of Manpower
Utilization
Headquarters, Marine Corps (Code MPU)
MCB (Building 2009
Quantico, VA 22134

1 Dr. A. L. Slafkosky
Scientific Advisor (Code RD-1)
Headquarters, U. S. Marine Corps
Washington, DC 20380

1 Chief, Academic Department
Education Center
Marine Corps Development and
Education Command
Marine Corps Base
Quantico, VA 22134

1 Mr. E. A. Dover
2711 South Veitch Street
Arlington, VA 22206

Coast Guard

1 Mr. Joseph J. Cowan, Chief
Psychological Research Branch (G-P-
1/62)
U. S. Coast Guard Headquarters
Washington, DC 20590

Other DOD

1 Dr. Harold F. O'Neil, Jr.
Advanced Research Projects Agency
Cybernetics Technology, Rm. 625
1400 Wilson Boulevard
Arlington, VA 22209

12 Defense Documentation Center
Cameron Station, Building 5
Alexandria, VA 22314
ATTN: TC

Other Government

1 Dr. Lorraine D. Eyde
Personnel Research and Development Center
U. S. Civil Service Commission
1900 E Street, N. W.
Washington, DC 20415

1 Dr. William Gorham, Director
Personnel Research and Development Center
U. S. Civil Service Commission
1900 E Street, N. W.
Washington, DC 20415

1 Dr. Vern Urry
Personnel Research and Development Center
U. S. Civil Service Commission
1900 E Street, N. W.
Washington, DC 20415

1 Dr. Harold T. Yahr
Personnel Research and Development Center
U. S. Civil Service Commission
1900 E Street, N. W.
Washington, DC 20415

1 Dr. Andrew R. Molnar
Technical Innovations in Education Group
National Science Foundation
1800 G Street, N. W.
Washington, DC 20550

1 U. S. Civil Service Commission
Federal Office Building
Chicago Regional Staff Division
Regional Psychologist
230 South Dearborn Street
Chicago, IL 60604
ATTN: C. S. Winiewicz

1 Dr. Carl Frederiksen
Learning Division, Basic Skills Group
National Institute of Education
1200 19th Street, N. W.
Washington, DC 20208

Miscellaneous

1 Dr. Scarvia B. Anderson
Educational Testing Service
17 Executive Park Drive, N. E.
Atlanta, GA 30329

1 Mr. Samuel Ball
Educational Testing Service
Princeton, NJ 08540

1 Dr. Gerald V. Barrett
University of Akron
Department of Psychology
Akron, OH 44325

1 Dr. Kenneth E. Clark
University of Rochester
College of Arts and Sciences
River Campus Station
Rochester, NY 14627

1 Dr. John J. Collins
Vice President
Essex Corporation
6305 Caminito Estrellado
San Diego, CA 92120

1 Dr. Rene V. Dawis
University of Minnesota
Department of Psychology
Minneapolis, MN 55455

1 Dr. Marvin D. Dunnette
University of Minnesota
Department of Psychology
Minneapolis, MN 55455

1 ERIC
Processing and Reference Facility
4833 Rugby Avenue
Bethesda, MD 20014

1 Major I. N. Evonic
Canadian Forces Personnel Applied Research Unit
1107 Avenue Road
Toronto, Ontario, CANADA

1 Dr. Victor Fields
Montgomery College
Department of Psychology
Rockville, MD 20850

1 Dr. Edwin A. Fleishman
Visiting Professor
University of California
Graduate School of Administration
Irvine, CA 92664

1 Dr. John R. Frederiksen
Bolt, Beranek and Newman, Inc.
50 Moulton Street
Cambridge, MA 02138

1 Dr. Robert Glaser, Co-Director
University of Pittsburgh
3939 O'Hara Street
Pittsburgh, PA 15213

1 Dr. Richard S. Hatch
Decision Systems Associates, Inc.
5640 Nicholson Lane
Rockville, MD 20852

1 Dr. M. D. Havron
Human Sciences Research, Inc.
7710 Old Spring House Road
West Gate Industrial Park
McLean, VA 22101

1 HumRRO Central Division
400 Plaza Building
Pace Boulevard at Fairfield Drive
Pensacola, FL 32505

1 HumRRO/Western Division
27857 Berwick Drive
Carmel, CA 93921
ATTN: Library

1 Dr. David Klahr
Carnegie-Mellon University
Department of Psychology
Pittsburgh, PA 15213

1 Dr. Alma E. Lantz
University of Denver
Denver Research Institute
Industrial Economics Division
Denver, CO 80210

1 Dr. Frederick M. Lord
Educational Testing Service
Princeton, NJ 08540

1 Dr. Robert R. Mackie
Human Factors Research, Inc.
6780 Corton Drive
Santa Barbara Research Park
Goleta, CA 93017

1 Dr. William C. Mann
University of Southern California
Information Sciences Institute
4676 Admiralty Way
Marina Del Rey, CA 90291

1 Mr. Edmond Marks
315 Old Main
Pennsylvania State University
University Park, PA 16802

1 Richard T. Mowday
College of Business Administration
University of Nebraska, Lincoln
Lincoln, NE 68588

1 Dr. Leo Munday, Vice-President
American College Testing Program
P. O. Box 168
Iowa City, IA 52240

1 Mr. Luigi Petrullo
2431 North Edgewood Street
Arlington, VA 22217

1 Dr. Steven M. Pine
University of Minnesota
Department of Psychology
Minneapolis, MN 55455

1 Dr. Diane M. Ramsey-Klee
R-K Research & System Design
3947 Ridgemont Drive
Malibu, CA 90265

1 Dr. Joseph W. Rigney
University of Southern California
Behavioral Technology Laboratories
3717 South Grand
Los Angeles, CA 90007

1 Dr. Andrew M. Rose
American Institutes for Research
3301 New Mexico Avenue, N. W.
Washington, DC 20016

1 Dr. George E. Rowland
Rowland and Company, Inc.
P. O. Box 61
Haddonfield, NJ 08033

1 Dr. Benjamin Schneider
University of Psychology
Department of Psychology
College Park, MD 20742

1 Dr. Lyle Schoenfeldt
Department of Psychology
University of Georgia
Athens, Georgia 30602

1 Dr. Arthur I. Siegel
Applied Psychological Services
404 East Lancaster Avenue
Wayne, PA 19087

1 Dr. Henry P. Sims, Jr.
Room 630 - Business
Indiana University
Bloomington, IN 47401

1 Dr. C. Harold Stone
1428 Virginia Avenue
Glendale, CA 91202

1 Dr. Patrick Suppes, Director
Institute for Mathematical Studies
in the Social Sciences
Stanford University
Stanford, CA 94305

1 Dr. Sigmund Tobias
PH.D Programs in Education
Graduate Center
City University of New York
33 West 42nd Street
New York, NY 10036

1 Dr. David J. Weiss
University of Minnesota
Department of Psychology
N660 Elliott Hall
Minneapolis, MN 55455

1 Dr. K. Wescourt
Stanford University
Institute for Mathematical Studies
in the Social Sciences
Stanford, CA 94305

1 Dr. Anita West
Denver Research Institute
University of Denver
Denver, CO 80210

1 Mr. George Wheaton
American Institutes for Research
3301 New Mexico Avenue, N.W.
Washington, DC 20016