INCOMPLETE ORDERS AND COMPUTERIZED TESTING

NORMAN CLIFF[1]
University of Southern California

A computerized adaptive testing system has three main aspects, and consequently it can differ in three main ways from a noncomputer system. First, there is the test item. Full utilization of a computer allows an enormous broadening in the type of problem that can be presented to the individual. Typing out objective questions to him is the most obvious thing to do, but it is far from the only thing, and is perhaps far from the best thing. There is perhaps even a greater extension of the possible types of examinee response, as we can see not only from what is described here but by borrowing from CAI techniques. Moreover, we can easily incorporate speed of response into the scoring; we can determine not only whether the person can give the answer, but whether he can give it in ten seconds. But the greatest difference between computerized adaptive testing and ordinary testing is in the extent and nature of the decision process that goes on between items.

It is with the latter aspect that I will be concerned here today; the approach suggested here is quite different conceptually than others such as the branching and the Bayesian methods, so the paper will trace its origins. Tests try to order persons, so we will first consider the basic nature of orders and then how orders can be constructed from incomplete data. Testing will be shown to be a type of ordering process which utilizes incomplete data; computerized adaptive testing develops orders from highly incomplete data. We will give a simple example of how a computer program based on these concepts works. Finally, some of the ways in which these concepts form the basis for a test theory will be suggested.

Our approach to a model for computerized testing has its origins in quite a different area, computer-interactive judgment methods. In order to demonstrate the relation between testing and ordering, let us consider for a moment a simple order. A simple order is defined, and please let me use quite informal language, as a set whose members display a relation between elements which demonstrates asymmetry and transitivity. Now what that means is that, if we have a matrix which records the existence of the relation as a 1, or its non-existence as a 0, between a pair of elements of the set, the matrix must display the triangular form shown in the first figure. Paired comparisons judgments of some stimulus property of course often display a close approximation to this form. For example, suppose we used the five indicated letters, presented them in pairs, and asked a child which came first in the alphabet. Then we record his judgment as a 1 if he responds that the row letter comes before the column letter and a 0 if he says the reverse. If he knew the order of the alphabet, then the data would be as shown.

	v	w	x	y	z
v	–	1	1	1	1
w	0	–	1	1	1
x	0	0	–	1	1
y	0	0	0	–	1
z	0	0	0	0	–

Fig. 1. Complete adjacency matrix for a simple order showing transitivity and asymmetry.

An interesting property of such paired comparisons matrices is that they need not be complete. Suppose we do not ask about all pairs, but do assume that the data is asymmetric and transitive. Then we may be able to complete the matrix by performing matrix algebra on the elements which we do have. This is illustrated in the second set of figures. The lefthand one shows an incomplete dominance matrix, one which incidentally would typically be found by the kind of interactive ordering program we developed, and the right one shows that matrix multiplied by itself. We see that in this instance the square of the obtained matrix shows exactly the same triangular form as the complete matrix in Fig. 1. Actually, the data matrix could be even more incomplete than this one and still yield a complete order. The *necessary* part of the matrix is the supradiagonal chain of ones which corresponds to the judgments concerning the letters which are next to each other in the alphabet. As long as we have these, then the matrix can be completed; we just have to raise it to a high enough power. Of course, when dealing with human judgments with their inconsistency, we have to build in some safeguards and redundancy in the process.

The reason for going through that exercise is that the model we propose for computerized testing is exactly the

	v	w	x	y	z
v	–	1		1	1
w	0	–	1		1
x		0	–	1	
y	0		0	–	1
z	0	0		0	–

	v	w	x	y	z
v	–		1		1
w		–		1	
x	0		–		1
y		0		–	
z	0		0		–

	v	w	x	y	z
v	–	1	1	1	2
w	0	–	1	1	1
x	0	0	–	1	1
y	0	0	0	–	1
z	0	0	0	0	–

Fig. 2. Sufficient adjacency matrix A_t, its square A_t^2 and the sum $A_t + A_t^2$, showing that the latter has the same qualitative form as A.

[1] Preparation of this paper was supported in part by the Office of Naval Research, Contract No. 150-373.

same! We say that tests order people. In what sense is that so? In what sense is the relation between people one which is asymmetric and transitive? It is superficially obvious that if examinees are given different scores, then the relation between the scores is asymmetric and transitive. That is just a property of numbers, in fact the one which served as a model for ordering in the first place. But it is a property which is just as true of the testees' zip-codes, or their social security numbers, or their football jersey numbers, as it is of their test scores. What is it about test scores that makes the order empirically meaningful rather than arbitrary?

Test scores start out from binary relations between people and items. How is it that we are allowed to derive from such relations numbers which give us an order of people, in the same sense that we can assign numbers to stimuli that give their order? Where is the asymmetric, transitive relation?

A long time ago, Louis Guttman gave part of the answer (Guttman, 1941). He said that items order persons if the score matrix displays the form we have come to call the Guttman scale, but should more fairly call the Guttman-Loevinger scale since she invented an almost identical concept and developed it in a superior way (Loevinger, 1947). But Guttman's answer is not completely satisfactory to the formalist. The score matrix is rectangular, not square; item responses are defined as right or wrong by fiat and have no chance to be other then asymmetric. The transitivity of a Guttman scale is indirect.

The most important part of the answer to the questions concerning the legitimacy of items as orderers of persons lies in the realization that the score matrix is only part of a larger matrix of relations. The relations matrix is really items-plus-persons by items-plus-persons, not just items by persons. We think of the response of a person to an item as indicating a dominance relation between the person and the item. Habitually, we put a one in the score matrix if the person gets the item right and a zero if he gets it wrong. But that is because, being people, we identify with the persons dimension of the matrix. If instead we were items, in some through-the-looking-glass world, we would use the opposite notation, giving the *item* a one if the person got it wrong and a zero if the dumb thing allowed itself to be gotten right by the person.

	a	b	1	2	3
a			0	1	1
b			0	0	1
1	1	1			
2	0	1			
3	0	0			

S

	a	b	1	2	3
a	0	1			
b	0	0			
1			0	1	2
2			0	0	1
3			0	0	0

S^2

Fig. 3. Complete (showing rights and wrongs) score matrix S for two items a, b and three persons, 1, 2, 3 for scalable data; and S^2 showing item-item and person-person dominance.

Taking the point of view of neither items nor persons but rather of test theorists, we must take a less chauvinistic stance and play fair in our scorekeeping. The score matrix is expanded. In the expanded matrix, we give a one to the winner of the contest between item and person and a zero to the loser, regardless of which is which. Such a matrix is given at the left of Figure 3. In the lower left corner of the matrix we have the usual binary score matrix which shows which items were defeated by which persons. The matrix here is of the Guttman form. In the upper right we have the same matrix from the item point of view, giving a one each time an item defeats a person. Since the score matrix is complete here, the upper right matrix is the transposed complement of the lower right one.

There are two other sections of this expanded score matrix and these are left blank. These sections correspond to the item-item and person-person relations, which are not observed directly. In the case of pairwise judgments, we found above that an incomplete matrix could be completed by squaring the observed matrix. Let us do that in the present case. The result is shown in the right side of the figure. It is two triangular matrices, one for items and one for persons. Thus, treated in this formal fashion, we see that a GL scale does give two asymmetric transitive relations, one for items and one for persons. We will return to these two order matrices in another context.

We can put the two orders together. This is illustrated in Figure 4; the matrix on the left is simply the sum of the two matrices from Figure 3, that is $S + S^2$. The matrix on the right of Figure 4 contains exactly the same elements, but they have been rearranged, that is, pre- and postmultiplied by a permutation matrix P, into the order which is implied here, a *joint* order of persons and items, which is seen to in fact be a simple order because of the triangular, i.e., asymmetric and transitive form of the matrix. This answers those querulous questions about where the order is in the case of test data. If the data are a Guttman scale, then the score matrix, expanded and operated on in the manner indicated, does indeed define an order in the rather strict sense of the existence of a relation on a set, a relation which is transitive and asymmetric.

Let me say that for illustrative purposes here the matrix operations have been carried out in ordinary arithmetic.

	a	b	1	2	3
a	0	1	0	1	1
b	0	0	0	0	1
1	1	1	0	1	2
2	0	1	0	0	1
3	0	0	0	0	0

$S + S^2$

	1	a	2	b	3
1	–	1	1	1	2
a	0	–	1	1	1
2	0	0	–	1	1
b	0	0	0	–	1
3	0	0	0	0	–

$P(S + S^2)P$

Fig. 4. $S + S^2$ in its original segregated form (left) and reordered form (right), the latter showing qualitative asymmetry and transitivity like a simple order.

Because the relations are logical rather than arithmetic, we should have been doing the matrix multiplication with Boolean arithmetic. The only thing that changes in the present context is that all numbers greater than one in the matrices should be set equal to one.

So far, we have not referred directly to anything having to do with "computerized adaptive testing," but the relevance of the above theoretical sketch is quite direct. Just as the score matrix itself is a kind of incomplete matrix of dominance relations that can be completed by the powering operation, an even more incomplete set of relations is all that is really necessary to define the joint person-item order. If we happen to ask each person only the hardest item he can answer correctly and the easiest item he would miss, those 2n relations—actually, 2n-2 is enough—are sufficient to define the complete joint order of items and persons. This subset of relations can quite simply be shown to correspond to the relations between adjacent elements in the order, the supradiagonal string of ones we saw in the incomplete paired comparisons matrix of Fig. 2. In fact, if you look at the righthand matrix of Figure 4, the string of ones just above the diagonal there denotes exactly this set of item-person relations. In the 1975 *Bulletin* article (Cliff, 1975) I illustrated the way in which such a set of

	a	b	c	d	1	2	3	4	5
a					0	1	0*	0*	0*
b					0	0	1	0*	0*
c		0			0*	0	0	1	0*
d					0*	0*	0	0	1
1	1	1							
2	0	1	1			0			
3		0	1	1					
4			0	1					
5				0					

A

	a	b	c	d	1	2	3	4	5
a					0	1	1*	1*	
b					0	0	1	1*	1*
c			0		0*	0	0	1	1*
d					0*	0*	0	0	1
1	1	1	1	0*					
2	0	1	1	1*		0			
3	0*	0	1	1					
4	0*	0*	0	1					
5		0*	0*	0					

$A(A + I)^{(2)}$

	a	b	c	d	1	2	3	4	5
a	0	1	1	1					
b	0	0	1	1		0			
c	0	0	0	1					
d	0	0	0	0					
1					0	1	1	1	0
2					0	0	1	1	1
3		0			0	0	0	1	1
4					0	0	0	0	1
5					0	0	0	0	0

$A(A + I)^{(3)}$

	a	b	c	d	1	2	3	4	5
a					0	1	1*	1*	1*
b			0		0	0	1	1*	1*
c					0	0	0	1	1*
d					0	0	0	0	1
1	1	1	1*	1*					
2	0	1	1	1*					
3	0*	0	1	1		0			
4	0*	0*	0	1					
5	0*	0*	0*	0					

$A(A + I)^{(4)}$

Fig. 5. Illustration of completion by powering. Starred entries are derived by implication.

relations could be used to reconstruct the complete score matrix. That process is reproduced here in Figure 5 where the matrix powering is carried out.

Unfortunately, there is a problem; we do not know the right items to ask a person until after we have asked them. The routine by which the computer searches for the right items to ask is one of the two main aspects of the processing part of computerized adaptive testing, the other main aspect being how it damps out error. In our research, what we are doing is carrying over some principles which we have previously found to be effective in the paired comparisons ordering case.

The next set of figures illustrate the operation of a prototype program of the kind we have in mind, written by Jerry Kehoe. First, the program asks each person two items at random. The entries in the lefthand matrix of Figure 6 show the results of these preliminary rounds and the righthand one shows the powered matrix which contains the implications of these responses as well as the responses themselves. So far these are very few. The computer then decides which items to ask which persons next by seeing which are closest together in the order so far determined. This process of presentation, powering, and selection would go on for several rounds. The next figure shows the score matrix for an intermediate round on the left and the implications on the right. Now the powering process is having some effect. The next one shows the final score matrix on the left and the implications on the right where we see that not only has the score matrix been completed by implication but there are now complete simple orders of persons and items.

We incidentally do not have a name for this method. We would like to call it the Extended Transitivity System, or ETS, but those initials have been preempted.

You can see that the savings are not very great in this instance; each person must be asked most of the items. This impression is primarily a function of the size of the data matrix here. The savings are much, much greater with large matrices. An upper bound for the number of item-person relations that must be observed for n persons and x items is $\log_2(n + x)!$. For 200 persons and 200 items this number is about 2886. That means we would need to ask each person only 15 items to get the complete order; moreover, this upper bound is quite a generous one in the present instance, a couple fewer might well be sufficient.

Thus the method will work if the responses form a Guttman scale. It works surprisingly quickly and requires surprisingly little space in the computer, primarily because the programs take advantage of the binary nature of the data to store responses as single bits and then to carry out many of the calculations on whole words, that is, 32 elements at a time are processed in raising the matrix to the next power.

It is really no surprise that it works with errorless data. The crucial questions are how well will it work with the kind of inconsistent items and persons that the real world

	items							persons					
	a	b	c	d	e	f	g	1	2	3	4	5	6
a													
b								0					
c								0	0				
d											1		
e									0		0		
f										0		0	1
g										0		0	1
1		1	1										
2			1		1								
3						1	1						
4				0	1								
5						1	1						
6						0	0						

	items							persons					
	a	b	c	d	e	f	g	1	2	3	4	5	6
a													
b								0					
c								0	0				
d					1						1		
e				0					0		0		
f										0		0	1
g										0		0	1
1		1	1										1
2			1		1								
3						1	1						
4				0	1								
5						1	1						
6						0	0	0					

Fig. 6. (Left) Initial item responses matrix S, showing both person dominances and item dominances. Blank entries indicate item-person pairs not yet observed. (Right) $S + S^2$, showing the implied item-item and person-person dominances.

	items							persons					
	a	b	c	d	e	f	g	1	2	3	4	5	6
a								1					1
b								0		1			1
c								0	0		1		
d									0		1		
e									0		0	1	
f										0		0	
g										0		0	1
1	0	1	1										
2			1	1	1								
3		0				1	1						
4			0	0	1								
5					0	1	1						
6	0	0				0							

	items								persons				
	a	b	c	d	e	f	g	1	2	3	4	5	6
a		1	1		1*	1*	1*	1		1*	1*	1*	1
b	0					1	1	0		1			1
c	0				1	1*	1*	0	0		1	1*	1*
d					1	1*	1*		0		1	1*	1*
e	0*			0	0	1	1	0*	0		0	1	1*
f	0*	0	0*	0*	0			0*	0*	0	0*	0	1
g	0*	0	0*	0*	0			0*	0*	0	0*	0	
1	0	1	1		1*	1*	1*			1	1	1*	1
2			1	1	1*	1*	1*			1	1	1*	1*
3	0*	0				1	1	0	0				1
4	0*		0	0	1	1*	1*	0	0			1	1*
5	0*		0*	0*	0	1	1	0*	0*		0		1
6	0	0	0*	0*	0*	0		0	0*	0	0*	0	

Fig. 7. (Left) intermediate item response matrix S. (Right) $S + S^{(2)} + S^{(3)} + S^{(4)} + S^{(5)}$. Starred (*) entries are derived by indirect implication, i.e., from $S^{(3)}$, $S^{(4)}$, or $S^{(5)}$.

	items							persons					
	a	b	c	d	e	f	g	1	2	3	4	5	6
a								1	1				1
b								0	1	1	1		1
c								0	0	1	1		
d								0	0	0	1		
e									0	0	0	1	
f										0		0	1
g										0		0	0
1	0	1	1	1									
2	0	0	1	1	1								
3		0	0	1	1	1	1						
4		0	0	0	1								
5					0	1	1						
6	0	0				0	1						

	items								persons				
	a	b	c	d	e	f	g	1	2	3	4	5	6
a		1	1	1	1	1*	1	1	1	1*	1*	1*	1
b	0		1	1	1	1	1	0	1	1	1	1*	1
c	0	0		1	1	1	1	0	0	1	1	1*	1*
d	0	0	0		1	1*	1*	0	0	0	1	1*	1*
e	0	0	0	0		1	1	0*	0	0	0	1	1*
f	0*	0	0	0*	0		1	0*	0*	0	0*	0	1
g	0	0	0	0*	0	0		0*	0*	0	0*	0	0
1	0	1	1	1	1*	1*	1*		1	1	1	1*	1
2	0	0	1	1	1	1*	1*	0		1	1	1	1*
3	0*	0	0	1	1	1	1	0	0		1	1	1
4	0*	0	0	0	1	1*	1*	0	0	0		1	1*
5	0*	0*	0*	0*	0	1	1	0*	0	0	0		1
6	0	0	0*	0*	0*	0	1	0	0*	0	0*	0	

Fig. 8. (Left) Final response matrix S, showing 26 of the 42 item-person combinations which were used. (Right) $S + S^{(2)} + S^{(3)} + S^{(4)} + S^{(5)}$ with starred (*) elements indicating those entered by indirect implication.

faces us with, and what advantages does it offer over other approaches? The answer to the first question must await the opportunity to test it first with artificial stochastic data and then with real data. How well it will do in practice relative to the other approaches that have been reported and which we are hearing about during these two days must await even further data.

A priori, the methodology here appears to offer at least one potential advantage, the avoidance of extensive pretesting to determine item characteristics. Such pretesting presented problems, even to paper and pencil testing. There was the security problem, the question of comparability of populations, the differing contexts, the expense itself. In the computerized situation, these all become more acute. The present process avoids pretesting since items and persons are processed in parallel.

This method does require a substantial number of persons being tested simultaneously, however; but this is only initially true. Once a substantial set of person-item relations has been built up, additional persons can be processed individually as they appear, being fit into the previously determined order by means of their responses to the items. Under that mode of operation the amount of additional computer processing would be quite small.

It also seems to me that this way of thinking about tailored testing makes it easier to think of testing as integrated into a total personnel process. After all, it could be that the item selected for a person at a given point could be something like, "You have been assigned to welders' school. Come back when you have completed the course." The "item" in that case is successful completion of the course.

But to me, the most promising aspect of this method is theoretical. It furnishes the basis for a test theory which I think is more appropriate to the computerized testing context. If what is wanted from testing is an order of persons, and norms after all just tell the individuals' positions relative to some benchmark persons, then surely we want the order to be consistent and complete. How do you tell if the order is consistent and complete? You look at the person-person relation matrix and see if it is asymmetric and transitive. It is easy to think of indices which would reflect the degree to which that matrix has those properties. Indeed, I had intended to spend my time here today talking about them, but the results of our study are not quite ready for presentation yet. Such indices furnish analogues of the familiar Kuder-Richardson formulas which are central to basic test theory, and in fact are related to them in the case of complete data. They have the additional property of being readily generalizable to the incomplete or computer-adaptive case. Thus if we go about computerized testing in the way described here, we can at least have appropriate evaluational indices built into the system. Other tailored testing schemes rely on external information from traditional modes of testing to get their biserial correlations, item difficulties, reliabilities, and so on. Here, analogues of these indices will come out of the interactive process itself.

REFERENCES

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

Guttman, L. The quantification of a class of attributes: A theory and method of scale construction. In P. Horst (Ed.), *The prediction of personal adjustment.* New York: Social Science Research Council, 1941.

Loevinger, J. A systematic approach to the construction and evaluation of Tests of Ability. *Psychological Monographs,* 1947, 61 (4, Whole No. 285).