DOCUMENT RESUME

ED 205 543 TM 810 421

AUTHOR Tatsuoka, Kikumi K.; Tatsuoka, Maurice M.
TITLE Detection of Aberrant Response Patterns and Their Effect on Dimensionality.
INSTITUTION Illinois Univ., Urbana. Computer-Based Education Research Lab.
SPONS AGENCY Office of Naval Research, Arlington, Va. Personnel and Training Research Programs Office.
REPORT NO CERL-RR-80-4
PUB DATE Apr 80
CONTRACT N00014-79-C-0752
NOTE 54p.

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Error Patterns; *Goodness of Fit; Item Analysis; Mathematical Models; Test Construction; Test Reliability
IDENTIFIERS Individualized Consistency Index; *Pattern Conformity Index; *Unidimensional Scaling

ABSTRACT

An index measuring the degree to which a binary response pattern conforms to some baseline pattern was defined and named the Pattern Conformity Index (PCI). One way of conceptualizing what the PCI measures is the extent to which each individual's particular response pattern contributes to, or detracts from, the overall consistency found in the group's mode of responding. One use of the PCI consists of spotting anomalous response patterns that result from a student's problems. From this it is a short step to utilizing the PCI for identifying a subgroup of students for whom the given set of items approximately constitutes a unidimensionally scalable set. The duality between students and items then permits selection of a subset of items for further improving the unidimensionality. A measure of how constant an individual's response pattern remains for parallel subsets of items ocurring earlier and later in a test was developed and called the Individualized Consistency Index (ICI). (Author/BW)

Computer-based Education
Research Laboratory

University of Illinois Urbana Illinois

DETECTION OF ABERRANT RESPONSE PATTERNS AND THEIR EFFECT ON DIMENSIONALITY

KIKUMI K. TATSUOKA
MAURICE M. TATSUOKA

This research was sponsored by the Personnel and Training Research Program, Psychological Sciences Division, Office of Naval Research, under Contract No. N00014-79-C-0752. Contract Authority Identification Number NR 150-415.

COMPUTERIZED ADAPTIVE TESTING AND MEASUREMENT RESEARCH REPORT 80-4

APRIL 1980

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Research Report No. 80-4	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Detection of Aberrant Response Patterns and their Effect on Dimensionality		5. TYPE OF REPORT & PERIOD COVERED Nov. 21 '79 - Feb. 20 '80
		6. PERFORMING ORG. REPORT NUMBER CERL Report E-15
7. AUTHOR(s) Kikumi Tatsuoka & Maurice Tatsuoka		8. CONTRACT OR GRANT NUMBER(s) N00014-79-C-0752
9. PERFORMING ORGANIZATION NAME AND ADDRESS Computer-based Education Research Laboratory University of Illinois Urbana, IL 61801		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61153N; RR042-04 RR042-04-01; NR154-445
11. CONTROLLING OFFICE NAME AND ADDRESS Personnel and Training Research Laboratory Office of Naval Research (Code 458) Arlington, VA 22217		12. REPORT DATE April 1980
		13. NUMBER OF PAGES 45
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report)
		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)

18. SUPPLEMENTARY NOTES

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

error analysis, response patterns, integer operations, order analysis, consistency index, pattern conformity index, subset of unidimensionality

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

An index measuring the degree to which a binary response pattern conforms to some baseline pattern was defined and named the Pattern Conformity Index (PCI). By "baseline pattern" we mean a binary response vector with all the 0's preceding the 1's when the items are arranged in descending order of difficulty or in some other, purposefully defined order.

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

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

It was shown that PCI is related to Cliff's consistency index C in the following manner. When the items are arranged in ascending order of the proportions of individuals in a group who pass the items, and the PCI is computed for each individual, a certain weighted average of these PCI's yields an index which is a slight modification of Cliff's C.

One way of conceptualizing what the PCI measures is the extent to which each individual's particular response pattern contributes to, or detracts from, the overall consistency found in the group's mode of responding.

The foregoing observations make it clear that one use of the PCI consists in spotting anomalous response patterns that result from a student's problems, for example. From this it is a short step to utilizing the PCI for identifying a subgroup of students for whom the given set of items approximately constitutes a unidimensionally scalable set. The duality between students and items then permits selection of a subset of items for further improving the unidimensionality.

Very roughly speaking, the culling out of students and/or items to achieve unidimensionality by using the PCI proceeds as follows. Students whose response pattern are so anomalous (i.e., so atypical of the group) as to have negative PCI values are eliminated from the outset. The weighted average of the PCI's of the remaining members of the group (referred to above as resembling Cliff's C) is computed. Then, in a manner somewhat analogous to removing variables in the backward elimination method of stepwise multiple regression, students are successively removed from the group in such a way that at each step the PCI-weighted-average for the remaining group shows the largest increment from the previous value. A suitable stopping rule terminates the process before the group gets intolerably small in size. A computer routine for effecting this procedure was developed.

The PCI measures the degree to which an individual's response pattern resembles the <u>group's</u> modal response pattern. Sometimes, however, we need a measure of how constant an individual's response pattern remains for parallel subsets of items ocuurring earlier and later in a test. One reason for this is that students often switch their rules of operation -- either from one erroneous rule to another or from an erroneous to the correct rule -- as they proceed through a test. Thus, their response patterns tend to be inconsistent among one another while learning is taking place, but become more and more consistent as they reach mastery level. An index for measuring individual consistency was developed and called the <u>Individualized Consistency Index</u> (ICI).

ACKNOWLEDGEMENTS

We wish to acknowledge the services rendered by the following persons.

Bob Baillie, for designing and writing several computer programs - tehcnical, data processing, collection routines as well as computerized instructional lessons.

Roy Lipshutz and Wayne Wilson for their artwork.

ABSTRACT

An index measuring the degree to which a binary response pattern conforms to some baseline pattern was defined and named the Pattern Conformity Index (PCI). By "baseline pattern" we mean a binary response vector with all the 0's preceding the 1's when the items are arranged in descending order of difficulty or in some other, purposefully defined order.

It was shown the PCI is related to Cliff's consistency index C_t in the following manner. When the items are arranged in ascending order of the proportions of individuals in a group who pass the items and the PCI is computed for each individual, a certain weighted average of these PCI's yeilds an index which is a slight modification of Cliff's C.

One way of conceptualizing what the PCI measures is the extent to which each individual's particular response pattern contributes to, or detracts from the overall consistency found in the group s mode of responding.

The foregoing observations make it clear that one use of the PCI consists of spotting anomalous response patterns that result from a student's problems, for example. From this it is a short step to utilizing the PCI for identifying a subgroup of students for whom the given set of items approximately constitutes a unidimensionally scalable set. The duality between students and items then permits selection of a subset of items for further improving the unidimensionality.

Very roughly speaking, the culling out of students and/or items to achieve unidimensionality by using the PCI proceeds as follows. Students whose response pattern are so anaomalous (i.e., so atypical of the group) as to have negative PCI values are eliminated from the outset. The weighted average of the PCI's of the remaining members of the group (referred to above as resembling Cliff's C) is computed. Then, in a manner somewhat analogous to removing variables in the backward elimination method of stepwise multiple regression, students are successively removed from the group in such a way that at each step the PCI-weighted-average for the remaining group shows the largest increment from the previous value. A suitable stopping rule terminates the process before the group gets intolerably small in size. A computer routine for effecting this procudre was developed.

The PCI measures the degree to hich an individual's response pattern resembles the group's modal response pattern. Sometimes, however, we need a measure of how constant an individual's response pattern remains for parallel subsets of items ocurring earlier and later in a test. One reason for this is that students often switch their rules of operation -- either from one erroneous rule to another of from an erroneous to the correct rule -- as they proceed through a test. Thus, their response patterns tend to be inconsistent among one another while learning is taking place, but become more and more consistent as they reach mastery level. An index for measuring individual-consistency was developed and called the Individualized Consistency Index (ICI).

DETECTION OF ABERRANT RESPONSE PATTERNS AND THEIR EFFECT ON DIMENSIONALITY

Kikumi Tatsuoka & Maurice Tatsuoka

INTRODUCTION

It would seem trite to say that the possiblity of an examinee's getting correct answers for the wrong reasons always lurks behind dichotomously scored test items and threatens to destroy the validity of the test -- except for the fact that, until recently, this possibility has largely been ignored by psychometricians. Although scattered attempts have been made to give partial credit for partial knowledge, procedures for discrediting correct answers arrived at by incorrect means have largely been confined to the use of formulas for correction for guessing. This paucity may not be devastating for standardized ability tests, but is practically fatal in the context of achievement testing that is an integral part of the instructional process. There the test must serve the purpose of diagnosing what type of misconception exists, so that appropriate remedial instruction may be given. This calls for delving into the cognitive processes that are brought into play in solving problems, and trying to pinpoint just where the examinee went astray, even when the correct answer was fortuitously produced.

This type of testing was pioneered by Brown & Burton (1978), whose celebrated BUGGY is essentially an adaptive diagnostic testing system which utilizes network theory for routing examinees through a set of problems in the addition and subtraction of positive integers.

The branching is such that each successive problem serves to narrow down the scope of "hypotheses" as to the type(s) of misconception held by the examinee until finally a unique diagnosis is made.

Tatsuoka et al. (1980) developed a diagnostic testing system which differed from BUGGY in that the test was not adaptive but "conventional", i.e., linear. The test was constructed for use in conjunction with lessons in the addition and subtraction of signed numbers (positive and negative integers) for eighth grade students and consisted of four parallel subtests of 16 items each. A system of error vectors was developed for diagnosing the type(s) of error committed.

Crucial to this system of error diagnosis is the ability to tell whether and to what extent a response pattern is "typical" or "consistent". We may speak of consistency with respect to either the average response pattern of a group or an individual's own response pattern over time. To measure consistency in these two senses, two related but distinct indices are developed in this paper. They are called the "Norm Conformity Index" (NCI) and "Individual Consistency Index" (ICI), respectively.

It is shown that a certain weighted average of the NCI's of the members of a group yields one of Cliff's (1977) group consistency indices, C_{t1}. The higher the value of C_{t1}, the closer the group data set is to being unidimensional in the sense of forming a Guttman scale. Response patterns produced by erroneous rules are usually quite different from the average response pattern. Hence, removing individuals

with low (usually negative) NCI values -- i.e., those with aberrant response patterns -- will yield a data set that is more nearly unidimensional.

The ICI, on the other hand, measures the degree to which an individual's response pattern remains invariant over time. Thus, for example, in the signed-number test consisting of four parallel subtests, the ICI indicates whether an examinee's response pattern changes markedly from one subset to the next or remains relatively stable. Low ICI values, indicating instability of response pattern, would suggest that the examinee was still in the early stages of learning, changing his/her method for solving equivalent problems from one wave to the next. A high ICI value, reflecting stability of response pattern, would signal the nearing of mastery or a learning plateau.

While the NCI and ICI can each serve useful purposes as suggested above and illustrated in detail below, examining them jointly opens up various diagnostic possibilities, as does the consideration of each of them in combination with the total test score.

NORM CONFORMITY INDEX

Cliff (1977) defined various consistency indices based on the notion of dominance and counterdominance relationships between pairs of items. Some of these are closely related to indices developed in the theory of scalability, originating in the 1930's. Although Cliff's indices are derived from the dominance matrix for the data set of an entire group, they can be expressed as weighted averages of the corresponding indices based on constituent subgroups of examinees. (Krus, 1975; Mokken, 1970; Yamamoto & Wise, 1980.) Nevertheless, it should be noted that these indices are measures of group consistency and do not represent individual examinees' consistency of responses.

Birenbaum & Tatsuoka (1980) demonstrated that individual response patterns offer powerful information for determining any erroneous rule of operation that a given examinee may have used in taking a test. Tatsuoka et al. (1980) developed a diagnostic system for identifying erroneous rules by generating "error vectors", each of whose binary elements represents the presence/absence of a specific "atomic" error. In this paper we develop an index that associates with each response-pattern vector a number between -1 and 1 (inclusive) representing the degree of concordance the vector shows with a Guttman vector of the same length (i.e., the same number of 1's) with the items arranged in some purposefully specified order. For instance, they may be arranged -- as they are in computing Cliff's indices -- in descending order of difficulty for the total group; or they may be arranged in any particular order that suits a given purpose.

It should be noted that group consistency in Cliff's sense is maximized when the items are ordered by difficulty. Any change of item ordering would result in a decrease in the value of any of Cliff's consistency indices. The value yielded by each of Cliff's formulas may, therefore, be regarded as a function of item order.

Consider a dataset consisting of just one person's response pattern row vector S. The dominance matrix for this response pattern is

(1) $$\bar{S}'S = N = (n_{ij}) \quad i,j = 1,2,\ldots,n \quad (= \text{number of items})$$

where $\bar{S}'$ is the transpose of the complement of S. By construction, $n_{ij} = 1$ when the individual gets item i wrong and item j right; otherwise $n_{ij} = 0$.

Of course if the ordering of the items in S is changed, the dominance matrix will also change. Consequently, the consistency index associated with response pattern S, defined as

(2) $$C_p = 2U_a/U - 1 \quad ,$$

where $U_a = \sum_i \sum_{j>i} n_{ij}$ (the sum of the above-diagonal elements of N),

and $U = \sum_j \sum_i n_{ij}$ (the sum of all the elements of N),

is a function of the item order, 0. To make this fact explicit, we write $C_p(0)$.

<u>Example 1</u>: Let S = (10110); then

$$N = \overline{S}'S = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} (10110) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \end{bmatrix}$$

Here $U_a = \sum_{i} \sum_{j>i} n_{ij} = 2$ and $U = 6$; hence from Equation (2),

$$C_p(0) = (2)(2)/6 - 1 = -1/3 \; .$$

<u>Example 2</u>: Let S = (00111), the Guttman Vector with three 1's.

Then

$$N = \overline{S}'S = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} (00111) = \begin{bmatrix} 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$U_a = 6$, $U = 6$; hence $C_p(0) = 12/6 - 1 = 1$.

<u>Example 3</u>: Let S = (11100), the "reversed Guttman vector". Then

$$N = \bar{S}'S = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} (11100) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \end{bmatrix}$$

Here $U_a = 0$, $U = 6$; hence $C_p(0) = 0/6 - 1 = -1$.

From the foregoing examples, the first two of the following properties of $C_p(0)$ may be inferred. The other properties are illustrated by further examples, and intuitive arguments are given to substantiate them. Their formal proofs are not difficult but tedious, and are therefore omitted.

<u>Property 1</u>: $-1 \leq C_p(0) \leq 1$

<u>Property 2</u>: If the order of items is reversed in S, the absolute value of $C_p(0)$ remains unchanged, but its sign is reversed.

Since $U = \sum_j \sum_i n_{ij} = \sum_j \sum_i (1-s_i)s_j$, it is invariant with respect to permutations of the elements of S. On the other hand, if the order of the elements of S is reversed, so that $s_i = s_{n-i+1}$, the U_a for the new dominance matrix will become $U'_a = \sum_i \sum_{j>i} n'_{ij} = \sum_{i=1}^{n} \sum_{j=i+1}^{n} (1-s_{n-i+1})s_{n-j+1}$, which can be shown to be equal to $U - U_a$.

Therefore $C_p(0) = 2(U-U_a)/U - 1$

$$= -2U_a/U + 1 = -C_p(0)$$

Property 3: The consistency index $C_p(0)$ associated with a $2 \times n$ data matrix, comprising two response-pattern vectors S_1 and S_2, is a weighted average of the $C_p(0)$'s associated with S_1 and S_2, respectively. #

If $S = \begin{bmatrix} S_1 \\ \hline S_2 \end{bmatrix}$, the consistency index for S is

$$C_p(0) = (\bar{S}'_1 \; \bar{S}'_2)\begin{bmatrix} S_1 \\ \hline S_2 \end{bmatrix} = \bar{S}'_1 S_1 + \bar{S}'_2 S_2 \; .$$

Therefore, if we let

$$U_k = \sum_j \sum_i n_{ij}^{(k)} \quad \text{for } k = 1,2$$

and

$$U_{ka} = \sum_i \sum_{j>i} n_{ij}^{(k)} \quad \text{for } k = 1,2 \; ,$$

it follows that the U and U_a for S are given by

$$U = U_1 + U_2$$

and

$$U_a = U_{1a} + U_{2a} \; .$$

Hence,

$$C_p(0) = \frac{2(U_{1a} + U_{2a})}{U_1 + U_2} - 1$$

$$= \frac{U_1}{U_1 + U_2} \frac{2U_{1a}}{U_1} + \frac{U_2}{U_1 + U_2} \frac{2U_{2a}}{U_2} - 1$$

$$= \frac{U_1}{U_1+U_2}\left(\frac{2U_{1a}}{U_1} - 1\right) + \frac{U_2}{U_1+U_2}\left(\frac{2U_{2a}}{U_2} - 1\right)$$

$$= W_1 C_p(0)_1 + W2C_p(0)_2$$

Remark: The two response patterns S_1 and S_2 may be either those of two individuals are of a single individual taking a set of items on two occasions (as in a repeated-measures design) or two parallel sets of items. In the first case the $C_p(0)$ associated with the 2xn data matrix would be an average $C_p(0)$ for the pair of individuals; in the second, it would be an average over two measurement occasions for one person.

By mathematical induction on Property 3, it follows that the $C_p(0)$ associated with an N x n data matrix

$$X = \begin{bmatrix} S_1 \\ \hline S_2 \\ \hline \vdots \\ \hline S_n \end{bmatrix}$$

is a weighted average of the $C_p(0)$'s associated with the individual response-pattern vectors $S_1, S_2, \ldots, S_N$. In particular, when the items are arranged in descending order of difficulty for the group comprising the N individuals, the $C_p(0)$ associated with X is one of Cliff's (1977) consistency indices, C_{t1}. For this particular ordering of items, we give the name "norm conformity index" to the $C_p(0)$'s

associated with the individual response patterns.

Definition: Norm Conformity Index, NCI

When the item ordering is in descending order of difficulty for a particular group (designated the "norm group"), the consistency index $C_p(O)$ associated with the individual's response pattern S is called the norm conformity index, denoted by NCI. Thus, NCI indicates the extent to which a response vector S approximates the Guttman vector (in which all the zeros are to the left of the 1's) with the same number of 1's, when the items are arranged in descending order of difficulty for the norm group.

With this definition, plus an expanded version of Property 3, we state the relationship between Cliff's consistency index C_{t1} and the NCI's for the individuals in the group as

Property 4: Cliff's consistency index C_{t1} is a weighted average of the NCI_k (k=1, 2, ..., N), with weights $w_k = U_k/U$; i.e.,

$$C_{t1} = \sum_{k=1}^{N} (U_k/U)\, NCI_k,$$

where

$$U_k = \sum_i \sum_{j>i} n_{ij}^{(k)}$$

and

$$U = \sum_{k=1}^{N} U_k .$$

Example 4: Let S_1 = (01011), S_2 = (00111) and S_3 = (00001) be the response-pattern vectors for three individuals. Then, by calculations similar to those shown in Examples 1, 2, and 3, we get

(upon writing NCI for $C_p(0)$)

$$U_{1a} = 5, \quad U_1 = 6, \quad NCI_1 = 2/3$$
$$U_{2a} = 6, \quad U_2 = 6, \quad NCI_2 = 1$$
$$U_{3a} = 4, \quad U_3 = 4, \quad NCI_3 = 1$$

Hence,

$$\begin{aligned} w_1 NCI_1 + w_2 NCI_2 + w_3 NCI_3 &= (6/16)(2/3) + (6/16)(1) + (4/16)(1) \\ &= 7/8 . \end{aligned}$$

On the other hand, with

$$X = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix},$$

$$\bar{X}'X = N = \begin{bmatrix} 0 & 1 & 1 & 2 & 3 \\ 0 & 0 & 1 & 1 & 2 \\ 0 & 1 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$U_a = 15, \; U = 16, \; NCI = 30/16 - 1 = 7/8,$$

thus verifying Property 4.

In the paragraph preceding Property 4, the order of the items was taken to be the order of difficulty for the group of which the individual was a member, for $C_p(0)$ to be called NCI. Actually,

as evident in the formal definition of NCI, the group need not be one to which the individual belongs. It can be any group which the researcher chooses for defining the baseline or "criterion order" of the items; hence our referring to it as the norm group, and the index as the <u>norm</u> conformity index. Thus, for example, we might be concerned with two groups of students with vastly different instructional backgrounds but similar abilities. It is then quite possible for the difficulties of various skills to be rather different in the two groups. We might take Group 1 as the norm group, thus arranging the items in descending order of difficulty for this group. We could compute NCI's for members of both Group 1 and Group 2 on the basis of this criterion order, and would probably find the mean NCI for the two groups to be significantly different. The following examples, based on real data, illustrate this.

<u>Example 5</u>: The seventh grade students of a junior high school were divided at random into two groups, which were given different lessons teaching signed-number operations. (Tatsuoka & Birenbaum, 1979). One sequence of lessons taught the operations by the Postman Stories approach (Davis, 1964) while the other used the number-line method.

After addition problems had been taught, a 52-item test including both addition and subtraction problems was administered to all students. A t-test showed no significant difference between the mean test score of the two groups, as indicated in Table 1. However, when NCI's were computed for all students, using the item-difficulty order in

Group 1 (the Postman-Stories group) as the baseline, there was a significant difference between the mean NCI of the two groups.

Insert Table 1 about here

The means of test scores and the Norm Conformity Index

	Group 1 (N = 67)	Group 2 (N = 62)	
Total Score			
mean	20.06	18.36	t = 1.190
SD	8.30	7.88	p > .05
NCI			
mean	.55	.45	t = 2.246
SD	.23	.23	p = .0264

Example 6: Tatsuoka & Birenbaum (1979) demonstrated that proactive inhibition affected the performance on tests in material learned through subsequent instructions. The response patterns of students who studied new lessons written by using a different conceptual framework from that of their previous instructions showed a significantly different performance pattern. By a cluster analysis, four groups among which response patterns are significantly different were identified. The NCI values for 91 students based on the order of tasks determined by the proportion correct in the total sample were calculated and analysis of variance was carried out. The F-value was significant at p = 0.05.

Insert Table 2 about here

Table 2

ANOVA of Norm Conformity Index for Four Groups With Different Instructional Backgrounds

Group	N	Mean of NCIs	F
1	34	0.18	3.62 with df = 3, 87
2	27	0.41	
3	20	0.35	
4	10	0.18	

Up to this point, the U_a and U in Eq. (2) defining $C_p(0)$ -- and hence NCI as a special case -- were defined in terms of the numbers of dominances and counter-dominances between item pairs in the dominance matrix N. We now show that U_a can be explicitly defined in terms of the proximity of a response vector S to a Guttman vector with the same number of 1's.

<u>Property 5</u>: Let S be a response-pattern vector of an examinee on an n-item test, N = $\bar{S}'S$ the associated dominance matrix, and

$$U_a = \sum_{i=1}^{n} \sum_{j>i} n_{ij} \quad .$$

Then U_a is also the number of transpositions required to get from S to the <u>reversed</u> Guttman vector (all 1s preceding the zeros).

Since $n_{ij} = (1 - s_i)s_j$, it follows that

$$U_a = \sum_i \sum_{j>i} (1 - s_i)s_j$$

is the number of ordered pairs (s_i, s_j) $[i<j]$ of elements of S such that $s_i = 0$ and $s_j = 1$. That is, if for each $s_i = 0$, we count the number of $s_j = 1$ to its right in S, then the sum of these numbers over the set of 0's in S is equal to U_a. But this is the same as the number of transpositions (interchanges of elements in <u>adjacent</u> (0,1) pairs) needed to transform S, step by step, into (1 1 ... 1 0 0 ... 0). Thus U_a is a measure of <u>remoteness</u> of S from the <u>reversed</u> Guttman vector, which is equivalently its <u>proximity</u> to the Guttman vector.

<u>Example 7</u>: Let S = (0 1 0 1 1). Then, S can be transformed into (1 1 1 0 0) by five successive transpositions:

$$(\underline{0\ 1}\ 0\ 1\ 1) \rightarrow (1\ 0\ \underline{0\ 1}\ 1) \rightarrow (1\ \underline{0\ 1}\ 0\ 1)$$
$$\rightarrow (1\ 1\ 0\ \underline{0\ 1}) \rightarrow (1\ 1\ \underline{0\ 1}\ 0) \rightarrow (1\ 1\ 1\ 0\ 0);$$

thus $U_a = 5$ by the present definition. On the other hand,

$$N = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

and $U_a = \sum_i \sum_{i>j} n_{ij} = 5$ by the earlier definition.

It may also be noted that, if we denote the number of 1's in the lower triangle of $\bar{S}'S$ by U_b, i.e.,

$$U_b = \sum_i \sum_{j>i} n_{ij} ,$$

then U_b is the numberof ordered pairs (s_j, S_i) $[j < i]$ of elements of S such that $s_i = 0$ and $s_j = 1$. Hence,

$$U = \sum_j \sum_i n_{ij} = U_a + U_b$$

is the number of pairs (s_i, s_j) with $s_i \neq s_j$ that can be formed from the elements of S. Thus, $U = x(n - x)$, where x is the number of 1's in S, or the test score earned by a person with response pattern S. Consequently, U_a/U and U_b/U are the proportions of (0,1) pairs and (1,0) pairs, respectively, among all possible ordered pairs (s_i, s_j) $[i < j]$ of unlike elements. When S is a Guttman vector (0 0 ...0 1 1 ... 1) $U_a = U$ and $U_b = 0$, because all ordered pairs of unlike elements are (0,1) pairs. Conversely, when S is a reversed Guttman vector (1 1 ... 1 0 0 ... 0), $U_a = 0$ and $U_b = U$. Hence U_a/U ranges from 0 to 1 as an increasing function of the degree to which S resembles (or is proximal to) a Guttman vector. Similarly, U_b/U measures the proximity of S to a reverse Guttman vector, or its remoteness from a Guttman vector. In fact U_a/U was denoted by U' and proposed as an index of "deviance" of score patterns by van der Flier (1977).

With the above redefinition of U_a and U, the sense in which NCI is a measure of the extent to which a response pattern approximates a Guttman vector should have become clearer.

$$NCI = 2U_a/U - 1$$

is a rescaling of U_a/U to have limits 1 and -1 instead of 1 and 0.

It should be noted that U_a/U, and hence also NCI, is undefined

for a person who has a test score of either 0 or n, since $U = x(n - x) = 0$ in both these cases. There are two ways (at least) in which to cope with this problem. The first is arbitrarily to set NCI = 1 when U = 0, which is analogous to setting 0! = 1. This is reasonable because U = 0 only for S = (0 0 ... 0) and S = (1 1 ... 1), both of which are Guttman vectors in the sense of having no zero to the right of any 1. The second solution is to redefine NCI itself as

$$(3) \quad NCI = 2(U_a + 1)/(U + 1) - 1 \quad ,$$

which will automatically make NCI = 1 for the all-correct and all-incorrect response patterns. Each of these solutions, however, gives rise to problems of its own, as shown in the discussion section below.

Property 6: Suppose S_1 and S_2 are two n-item response patterns with the same number x of 1's, and that S_2 results from S_1 by applying t successive transpositions. Then

$$C_p(S_2) = C_p(S_1) \pm 2t/x(n - x) \quad ,$$

where the + sign is taken when S_2 is closer to a Guttman vector than is S_1 and the - sign when the opposite is true. #

From Property 5 the U_a associated with a given response pattern S is the number of transpositions necessary for getting from S to the reversed Guttman vector with the same number of 1's. Hence, if t is the number of transpositions it takes to get from S_1 to S_2, it

follows that

$$U_{2a} = U_{1a} + t$$

if S_2 is farther from the reversed Guttman vector, i.e., closer to the Guttman vector, than is S_1, and

$$U_{2a} = U_{1a} - t$$

if the opposite is true. Consequently,

$$\begin{aligned} C_p(S_2) &= \frac{2U_{2a}}{U} - 1 = \frac{2(U_{1a} + t)}{U} - 1 \\ &= \frac{2U_{1a}}{U} - 1 + \frac{2t}{U} \\ &= C_p(S_1) + \frac{2t}{x(n-x)} \end{aligned}$$

when S_2 is closer than S_1 to the Guttman vector. The sign preceding $2t/x(n-x)$ becomes $-$ when S_2 is farther than S_1 to the Guttman vector.

<u>Example 8</u>: Let $S_1 = (1\ 0\ 1\ 0\ 1\ 1)$ and $S_2 = (0\ 1\ 0\ 1\ 1\ 1)$. It takes two transpositions to get from S_1 to S_2:

$$(\underline{1\ 0}\ 1\ 0\ 1\ 1) \rightarrow (0\ 1\ \underline{1\ 0}\ 1\ 1) \rightarrow (0\ 1\ 0\ 1\ 1\ 1),$$

and S_2 is closer than S_1 to the Guttman vector $(0\ 0\ 1\ 1\ 1\ 1)$. Therefore, by Property 6 we should have

$$\begin{aligned} C_p(S_2) &= C_p(S_1) + (2)(2)/(4)(2) \\ &= C_p(S_1) + 1/2 . \end{aligned}$$

For the two response patterns, we have

$$U_{1a} = 5 \quad \text{and} \quad U_{2a} = 7,$$

so that

$$C_p(S_1) = (2)(5)/8 - 1 = 1/4$$

and

$$C_p(S_2) = (2)(7)/8 - 1 = 3/4 \quad ,$$

satisfying the above relation.

Property 7: The weights applied to individual NCI's in computing Cliff's consistency index C_{t1} (cf. Property 4) are invariant of changes in the baseline order of items. #

This is true because the weights

$$w_p = \frac{U_p}{\Sigma U_p}$$

depend only on $U_p = x_p(n - s_p)$, where x_p is the total score earned by person p -- i.e., on the number of 1's in response pattern S_p, and not on their positions.

It follows that NCI's associated with response patterns yielding scores close to n/2 get high weights while those corresponding to extreme scores get low weights. It is also seen that when the number of persons is large, each w_p is a fairly small positive number, while the NCI has a value between 1 and -1. Negative NCI's are an obstruction to having a large group consistency index, C_{t1}.

INDIVIDUAL CONSISTENCY INDEX

In the preceding section we defined, and described various properties of, an index which measures the extent to which an individual's response pattern "conforms" to that of a norm group. In some situations it is desirable to measure the extent to which an individual's response pattern remains unchanged or "consistent" over the passage of time. For example, it is reasonable to expect that, when a student is in the process of learning -- and hence presumably modifying the cognitive processes by which he/she attempts to solve problems -- his/her pattern of responses on successive sets of similar items will change considerably from one set to the next. When the student approaches mastery or a "learning plateau", his/her response pattern will probably remain relatively consistent from one set to the next. To define an index, called the Individual Consistency Index (ICI), that will serve to measure the degree of consistency (or stability) of an individual's response pattern over time, and to investigate its properties, are the purposes of this section. In the interest of clarity and ease of exposition, we embed our discussions in the context of an actual experimental study.

A 64-item, signed-number test was administered to 153 seventh graders at a junior high school. The test comprised 16 different tasks being tested by four parallel items each. The items were arranged so that four parallel subtests were successively given to each testee. Within each 16-item subtest, the order of items was randomized. Thus, for each examinee there are four response-pattern vectors with 16

elements each. The Individual Consistency Index (ICI) is defined on these four replications. We shall come back to this test later, but we first introduce ICI by a simpler example. Suppose a person took four parallel tests A, B, C, D with seven items each, and that his/her response patterns were as shown in the second column of Table 3. Also shown in this table are $U = x(7 - x)$ for each response pattern, the number U_a of transpositions needed to transform each response pattern into a reverse Guttman vector, the $C_p(0)$ for each response pattern, and the weight to be applied to each $C_p(0)$ for getting an overall index.

Table 3

Four response patterns and various quantities associated with them.

parallel test #(j)	Response Pattern	U_j	U_{ja}	$C_p(0)_j$	w_j
1	(1010010)	12	4	-.333	.286
2	(0010010)	10	6	.200	.238
3	(1000010)	10	4	-.200	.238
4	(1000010)	10	4	-.200	.238

The weighted average

$$\sum_{j=1}^{4} w_j \, C_p(0)_j = -.143$$

would be Cliff's consistency index C_{t1} if the four response patterns of Table 2 were those of four individuals and if the items had been arranged in their difficulty order for the group. Let us rearrange the items (or rather the sets of parallel items) in their order of difficulty for the

for the person, which is (2,4,5,7,3,1,6). The response patterns and other quantities occurring in Table 3 now become as shown in Table 4, which also has a new column showing the number of transpositions t_j necessary to get from the jth response pattern in Table 3 to the new one here.

Table 4

Response patterns resulting from those in Table 2 by arranging the items in difficulty order, and various associated quantities.

Parallel test # (j)	Response Pattern	t_j	U_j	U'_{ja}	$C_p(0')_j$	w_j
1	(0000111)	8	12	12	1.0	.286
2	(0000101)	3	10	9	.8	.238
3	(0000011)	6	10	10	1.0	.238
4	(0000011)	6	10	10	1.0	.238

Note that the new $C_p(0)$ for each response pattern satisfies Property 6:

$$C_p(0')_j = C_p(0)_j + 2t_j/U_j$$

The weighted average of the new $C_p(0)$ values is

$$\sum_{j=1}^{4} w_j C_p(0')_j = .9524 .$$

This is what we call the Individual Conformity Index, ICI. We may state its definition as follows.

<u>Definition</u>: Individual Conformity Index (ICI). Given a set of response patterns shown by a single individual on a set of parallel tests, we arrange the parallel items in their overall order of difficulty

for the individual and compute the $C_p(0)$ for each response pattern thus modified. If we now form a weighted average of these $C_p(0)$'s as though we were computing Cliff's C_{t1} in accordance with Property 4, the result is the ICI.

Remark: Note that ICI is an attribute of a single individual, not of a group as is Cliff's consistency index. ICI differs also from NCI in that the latter (also an individual attribute) depends on the baseline order of items, i.e., the difficulty order in some group specified as the norm group, whereas ICI is computed for an individual with no reference to any group. Rather, ICI requires that the individual in question has taken two or more parallel tests, and measures the consistency of his/her response patterns across these parallel tests.

Property 8: Since the parallel items are arranged in their order of difficulty for the individual in question when ICI is computed, while they are arranged in their order of difficulty for a norm group when NCI is computed, it follows that

$$ICI \geq NCI$$

for each examinee.

APPLICATION TO ERROR ANALYSIS: I

Birenbaum & Tatsuoka (1980) found that 1-0 scoring based simply on right or wrong answers caused serious problems when erroneous rules of signed-number operations were used by many examinees. The point is that many erroneous rules can lead to correct answers in many items.

To highlight the extent of the problem, Tatsuoka et al. (1980) developed an almost exhaustive set of 72 erroneous rules for doing addition and subtraction of pairs of signed numbers, and enumerated the number of correct answers that would result from consistently using each incorrect rule for a set of 16 items. The resulting histogram is shown in Figure 1, where it can be seen that, in an extreme case, 12 out of the 16 items could be answered correctly by an erroneous rule of operation.

Using real data from a 64-item test consisting of four parallel subtests of 16 items each, Birenbaum & Tatsuoka first did a principal components analysis on the original data -- with the items scored 1 or 0 in the usual manner. Next, the data were modified by giving a score of 0 when an item was correctly answered presumably by use of an erroneous rule, and another principal components analysis was done. (Details of how it was judged that a correct answer was arrived at by an incorrect rule are given in Birenbaum & Tatsuoka, 1980.) The change between the two analyses was dramatic. The dimensionality of the data became much more clearcut with the modified data. The item-total correlations became much higher, while the means of the 16 tasks (each represented by four parallel items) did not change significantly.

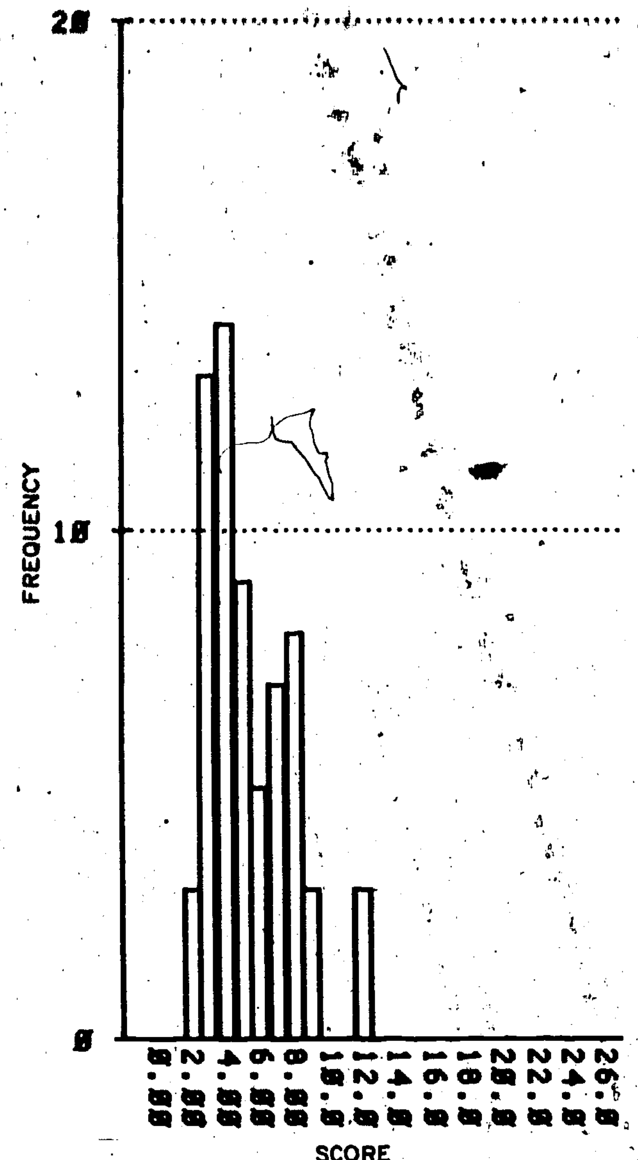

Figure 1. Histogram of total scores generated by erroneous rules of operation.

The above phenomenon suggests why some achievement tests cannot be treated as unidimensional even though the items are taken from a single content domain. The fact that correct answers can be obtained by erroneous rules apparently makes for a chaotic, multidimensional domain (Tatsuoka & Birenbaum, 1979), which is "cleaned up" by the rescoring. Brown & Burton (1978) warned of the same problem, namely that wrong rules can yield the correct answers in some items involving addition and subtraction of positive integers.

The indices developed in this paper, NCI and ICI, are useful for detecting erroneous rules that are consistently used by an examinee or a group of examinees. This capability is useful not only in the teaching process, for diagnosing a student's problem, but also gives some leads to addressing some psychometric issues such as the dimensionality of achievement tests.

Table 5 shows a 2 x 2 contingency table based on combinations of high and low NCI and ICI values, with a characterization of the status of students in each cell, dependent also on the score earned.

Table 5

Types of students with high and low NCI, ICI and score.

NCI \ ICI	Low	High
High	There should be few students in this cell (none if the cutting points for ICI and NCI are the same, since ICI $\geq$ NCI always).	If score is high, all is well. If score is low, student has a serious misconception (consistently uses an incorrect rule) which, however, leads to correct answers to easy items and wrong answers to hard items.
Low	The errors are probably random.	If score is high, student is merely getting <u>a few</u> of the easy items wrong. If score is low, student is getting <u>many</u> of the easy items wrong. The response pattern is strange, and a serious problem exists.

Example 9: The example described at the beginning of the section on the Individual Consistency Index was thoroughly analyzed in Technical Report 80-1 (Birenbaum & Tatsuoka, 1980) with respect to error analyses. We call this data the November data hereafter. There are 16 different erroneous rules diagnosed in the report. Table 6 shows the response patterns and NCI and ICI values for three students. Student 1 performed all addition problems correctly but he failed to change the sign of the subtrahend when he converted subtraction problems to addition problems. Student 2 always added the two numbers and took the sign of larger number for her answers. She failed to discriminate subtraction problems from addition problems and applied this erroneous rule consistently to all 16 tasks. Student 3 achieved fairly well but he occasionally mistyped or made careless mistakes. Determining the rules of operation, both right and wrong, is discussed in detail in the technical report 80-2 (Tatsuoka et al., 1980).

Table 6

The response patterns, NCI, ICI and scores of three students.

	Responses to four parallel forms within the 64-item test		
Task No.* Example	Student 1	Student 2	Student 3
6 6 + 4	1111 (+10)	1111 (10)	1111
15 -6 + 4	1111 (-2)	0000 (-10)	1111
3 12 + -3	1111 (9)	0000 (+15)	1111
5 -3 + 12	1011 (9)	0000 (+15)	1111
10 -14 + -5	1111 (-19)	1111 (-19)	1111
11 3 + -5	1111 (-2)	0000 (-8)	1111
14 -5 + -7	1111 (-14)	1111 (-12)	1111
7 8 - 6	0000 (+14)	0000 (+14)	1111
8 -16 - (-7)	0000 (-23)	0000 (-23)	1111
16 2 - 11	0000 (+13)	0000 (+13)	0111
13 -3 - +12	0000 (+9)	0000 (+15)	1111
1 -6 - (-8)	0000 (-14)	0000 (-14)	1111
12 9 - (-7)	0000 (+2)	1111 (+16)	0011
4 1 - (-10)	0000 (-9)	0000 (-11)	1010
2 -7 - 9	0000 (+2)	0000 (+16)	1111
9 -12 - 3	0000 (-9)	1111 (-15)	0111
NCI	0.9759	-0.2560	0.7073
ICI	1.0000	1.0000	.9268
Score	27	20	58

*The tasks are ordered by their overall difficulties over four parallel forms.

Table 6(a) is the 2 x 2 contingency table with .90 as the dividing point for ICI and .60 for NCI in the subsample of 75 students who earned scores of 53 or higher, while Table 6(b) is the corresponding table for 47 students with scores of 52 or lower. The dividing point, 52, for scores was chosen because -- as shown in Figure 1 -- 12 out of each subtest of 16 items could conceivably be answered by consistent use of an erroneous rule; 13 is the smallest number of items that cannot be got correct in this way, which corresponds to 52 out of the entire test of four parallel subtests of 16 items each. Hence it seems reasonable to regard 52 or less as "low scores".

Table 7

Two-way classifications based on ICI ≥ or < .90
and NCI ≥ or < .60 among students with
(a) scores ≥ 53; (b) scores ≤ 52.

NCI \ ICI	.90	.90	
.60	8	18	26
.60	26	23	49
	34	41	75

(a)

NCI \ ICI	.90	.90	
.60	3	27	30
.60	12	5	17
	15	32	47

(b)

Let us see what we can say about the performances of the students represented in the November data, from the contingency tables of Table 7, in light of the characterizations given in Table 5 and with the three students' response patterns in Table 6 to guide us to some extent. Note

first that, despite the very high cut-off point of .90 for the "high ICI" category, substantially more than one-half (73 out of 122) of the students have high ICI values. This reflects the fact that the examinees were eighth graders who had already received fairly extensive instruction in signed-number operations and hence a relatively large proportion of them showed stable response patterns over the four parallel subtests: they had already approached mastery or learning plateaus -- the latter being more likely in this case in view of the fact that only 75 (or 61.5%) of them had scores over 52 out of 64. As expected, very few students (11 out of 122) had low ICIs combined with high NCIs. Many more had low-ICI, low-NCI combinations; these are students who made more or less "random" (or at least non-systematic) errors but who nevertheless made relatively more errors among items that were easy for the group as a whole. It is reasonable that about 70% of the low scorers who had low NCIs fell in this category, while only 53% of the high scorers with low NCIs did.

Returning to the high-ICI group, members of which the three students represented in Table 6 are different kinds of examples, the high-ICI, high-NCI students with high scores are the "problem-free" types exemplified by Student 3. Unfortunately there are only 18 such students while there are 27 high-ICI, high-NCI students with low scores. Student 1 is an example of this type of student, and his response patterns corroborate the characterization in Table 5, that he has a serious misconception, but one which leads to correct answers (except for one probably careless error) to the easy items (addition) and always to wrong answers to the

hard items. In the high-ICI, low-NCI category, most students (23 out of 28) are high scorers. These students can easily be "remediated", for they are probably getting only a few easy items wrong. (It is an unfortunate fact that a few easy items missed can cause the NCI to become quite low.) The students who have the most serious problems are the high-ICI, low-NCI low scorers, of whom there are fortunately only five. Student 2 in Table 6 exemplifies this type, and her unusual response pattern (which remains perfectly consistent over the four parallel subtests) will take quite a bit of remedial instruction to rectify.

APPLICATION TO ERROR ANALYSIS: II

In the previous section, it was shown how the existence and seriousness of a student's misconception(s) could be determined by examining the |CI-NCI|-score combinations with each quantity dichotomized as high/low. Another concern we have in error analysis is to discover the extent to which a dataset has been "contaminated" -- in the sense of its having its unidimensionality destroyed -- by the consistent occurrence of erroneous rules of operation. This can be done by drawing a scatterplot of NCI against total score for the given dataset and comparing it with the corresponding scatterplot based on synthetic data generated by the 72 erroneous rules referred to at the beginning of the previous section. It is convenient also to draw the regression line of NCI on score for the synthetic data.

The procedure is illustrated in Figure 2, where the scatterplot of NCI against "task-mastery score*" for the November data (points represented by X's) is superimposed on the corresponding scatterplot for the artifically produced data generated by erroneous rules (points represented by o's) with the regression line shown. It can be seen at a glance that almost all of the real data points fall above the regression line for the synthetic data. In fact, a large majority of them even fall above the dashed line parallel to the regression line, which is drawn one S.E. above the latter.

*The "task-mastery score" is defined as follows: If a student gets at least three of the four parallel items testing a given task, his/her mastery score for that task is 1; otherwise it is 0.

We may conclude from the foregoing graphical inspection that the November data were fairly "clean". By contrast, the data obtained from the seventh graders tested after being taught signed-number operations via two PLATO sessions of about one hour each (see Example 5 above) were found to be highly "contaminated". Figure 1 shows the scatterplot of NCI against test score for one of these datasets, in which many points fall in the region occupied by the synthetic datapoints in Figure 2, suggesting that uses of erroneous rules abounded.

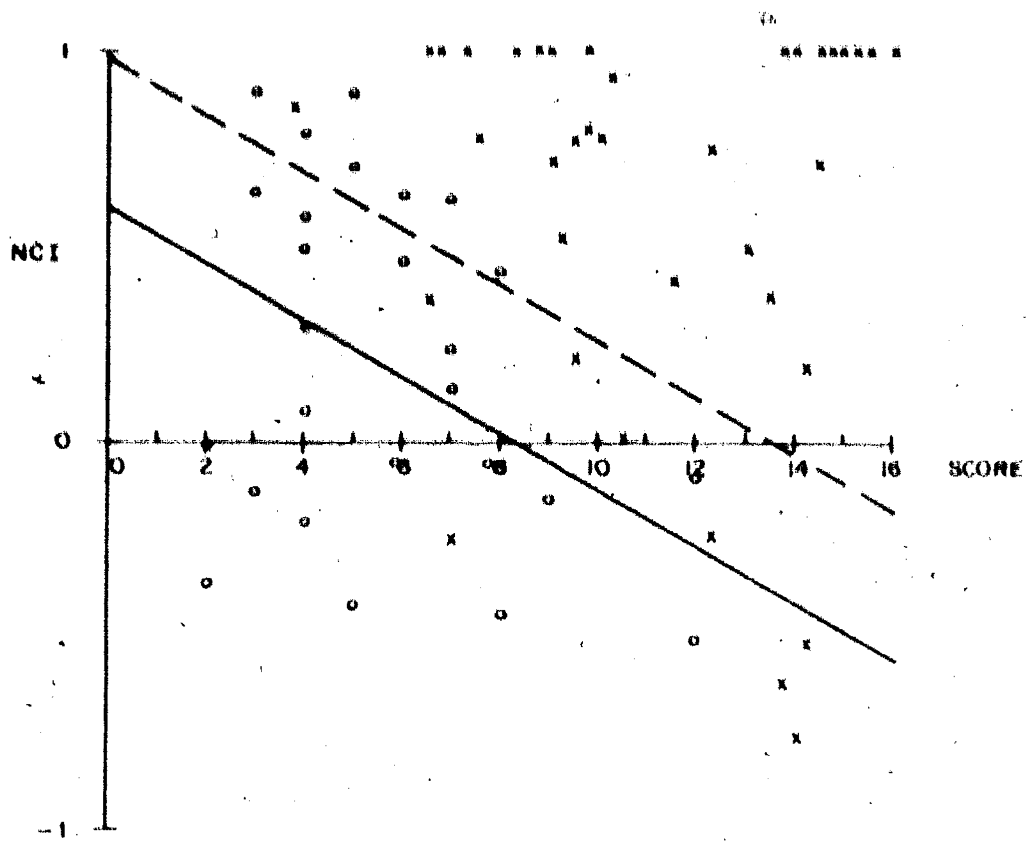

Figure 2. Scatterplot of NCI vs. total score for the November mastery score data, superimposed on that for 72 synthetic response patterns generated by erroneous rules.

x = real data (N = 127)

o = synthetic data (N = 72)

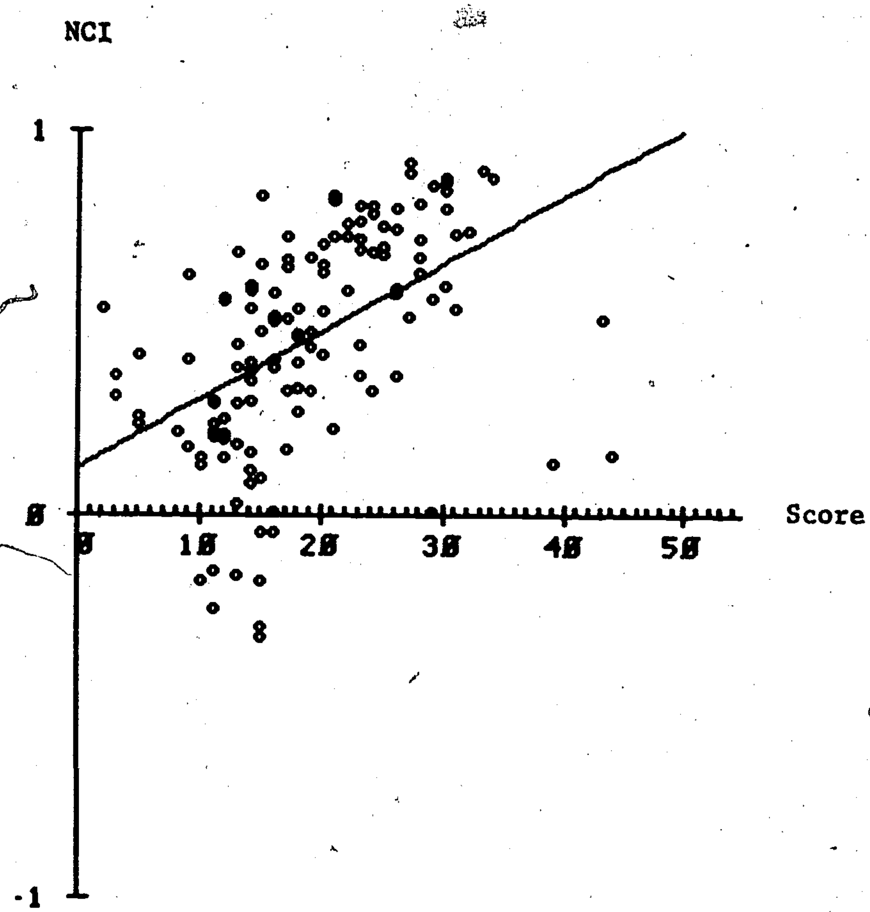

Figure 3. Scatterplot of NCI vs. total score for the January Postman Stories group posttest (N = 60)

EXTRACTING UNIDIMENSIONAL SUBSETS

Item Characteristic Curve (ICC) theories are useful and powerful test-theory models especially for applications in adaptive testing. If the test items are drawn from a single, unidimensional domain, logistic models are convenient for estimating the item-curve parameters. Tatsuoka (1980) examined the response patterns of students for whom ability estimates with known item parameters failed to converge by the maximum-likelihood method, and found them all to have low NCI values. Conversely, when the NCI is very small, the maximum-likelihood method often fails to yield a convergent estimate for the ability variable θ. Table 8 shows three such response patterns (along with one for which the θ estimate did converge) for the 48-item Stanford Vocabulary Test taken by the seventh graders in the January experiment. The item parameters for the three-parameter logistic model fitted to these data were estimated by LOGIST (Wood, Wingersky & Lord, 1976).

Table 8

One Convergent and Three Nonconvergent Response Patterns For Estimating θ by the MXL Method and Their NCI Values

	Response Patterns (48 items ordered roughly from easier to harder)	estimated θ	NCI	No. of iterations
1[b]	000010110000000010001100000000010000100000001001	-3.789	.0158	25[a]
3[b]	111100001000000000000000000100010000100110000000	-1.335	.2526	7
4	000000000000000000000000000000000000000111111111	-4.957	-1.0	25
5	000000000000000000000000000111111111100000000000	-25.00	-.4739	25

a Iterations were terminated at 25 tentatively, but the decrements of three cases exceeded .001.

b These two response patterns are taken from the real data and the other two are hypothetical response patterns.

The problem of non-convergence of maximum-likelihood estimation procedures for ICC models due to failure of the data to exhibit unidimensionality has been plaguing researchers for a long time. Mokken (1970) goes as far as to state that, although ICC theory has many valuable features, studies in sociology and political science are not quite ready to take advantage of the refined parameter estimation methods that it offers. To cope with this situation, he developed a techinque for extracting scalable subsets of items from a given dataset. Similar techniques have been developed in the fields of educational and psychological testing (Krus, 1977; Reynolds, 1976; Yamamoto & Wise, 1980). Theoretically, all these methods (which are based on order analysis) show a duality between items and examinees, and hence can in principle be used for extracting subsets of examinees as well as items in which unidimensionality will hold. In practice, however, most if not all of them would be quite ineficient for extracting examinee subsets because the dominance matrix for this purpose would be of order equal to the number of examinees instead of test items, and would thus be very large. We therefore present a new technique, based on the NCI, for extracting unidimensional examinee subsets that does not require the use of a dominance matrix as the starting point and hence is probably more efficient than those that do (although we have not yet made a formal comparison). Our technique is, of course, just as applicable for extracting unidimensional item subsets, but in that case its advantage, if any, over previous methods is probably negligible.

Recalling that Cliff's group consistency index C_{t1} is a weighted average of the NCI's of the members of the group (with the group itself defining the item-difficulty order), it is clear that individuals with negative NCI values are detrimental to the goal of getting a large C_{t1} value. We therefore remove any individual with a negative NCI from the group at the outset, before starting our extraction procedures. Let the NCI for the j-th member of the group thus purged be

$$NCI_j = \frac{2U_{ja}}{U_j} - 1 \quad .$$

Then, by Property 4 the overall consistency of the group, whose size we denote by N, is

$$C_p(S_N) = \sum_{j=1}^{N} w_j \; NCI_j \quad ,$$

where

$$w_j = U_j / \sum_{j=1}^{N} U_j \quad .$$

Now suppose we remove the k-th member of the group and denote the reduced data matrix by X_{N-1}. Then*,

*It is realized that removal of an individual from a group may conceivably alter the difficulty-order of the items, and hence the NCI's; but this is improbable, especially when the deleted individual has a small NCI, as we shall presently see he/she does.

$$C_p(X_{N-1}) = \sum_{j \neq k}^{N} w_j^{*} NCI_j$$

where

$$w_j^{*} = U_j / \sum_{j \neq k}^{N} U_j = U_j / \left(\sum_{j=1}^{N} U_j - U_k \right) .$$

The resulting change in overall consistency is

$$\Delta C_p = C_p(X_{N-1}) - C_p(X_N)$$

$$= \sum_{j \neq k} w_j^{*} NCI_j - \sum_{j=1} w_j NCI_j$$

$$= \left(\sum_{j=1}^{N} w_j^{*} NCI_j - w_k^{*} NCI_k \right) - \sum_{j=1}^{N} w_j NCI_j$$

$$= \sum_{j=1}^{N} (w_j^{*} - w_j) NCI_j - w_k^{*} NCI_k .$$

But

$$w_j^{*} - w_j = \frac{U_j}{\sum_{j=1}^{N} U_j - U_k} - \frac{U_j}{\sum_{j=1}^{N} U_j}$$

$$= U_j \frac{U_k}{(U_. - U_k) U_.} = \frac{U_k}{U_. - U_k} \cdot \frac{U_j}{U_.}$$

$$= \left(\frac{U_k}{U_. - U_k} \right) w_j ,$$

where U. is an abbreviation for $\sum_{j=1}^{N} U_j$.

Therefore

$$(4) \qquad \Delta C_p = \frac{U_k}{U. - U_k} \sum_{j=1}^{N} w_j NCI_j - w_k^* NCI_k$$

$$= \frac{U_k}{U. - U_k} \left[C_p(X_N) - NCI_k \right] \quad ,$$

since

$$w_k^* = U_k/(U. - U_k) \quad .$$

We thus see that, in order to make ΔC_p as large as possible -- that is, to have the removal of the k-th member result in as large an increase in C_p as possible -- two conditions must be satisfied; namely, NCI_k should be as small as possible, while U_k should be as large as possible. The first of these conditions is intuitively obvious. Since the overall C_p is a weighted average of the individual NCI's, elimination of the smallest NCI would be expected to increase the group C_p the most. However, since the latter is a weighted rather than a straight average of the NCI's, it is also necessary that the NCI to be eliminated have as high a weight as possible, namely, that the associated U_k be as large as possible. Recalling that

$$U_k = x_k(n-x_k) \quad ,$$

it follows that x_k should be as close as possible to n/2 in order for U_k to be large.

From a purely mathematical standpoint, the above optimizing condition for ΔC_p would require our actually computing ΔC_p for each potential individual to be removed, for there could be a tradeoff between NCI_k being small and U_k being large (i.e., $|x_k - n/2|$ being small). In practice, however, it is highly unlikely that a person with a small NCI would have a middling score that could yield a large U_k. That is, practically everyone with a small NCI would have a relatively extreme score, leading to a fairly small U_k. Hence, the smallness of NCI becomes the overriding consideration in selecting the individual to be removed. It therefore suffices to compute ΔC_p in accordance with Equation (4) for just those members of the group who have the few smallest values of NCI and select the one among them that yields the largest ΔC_p.

In the foregoing manner, we would successively remove the member of the remaining group that produces the largest increase in the overall C_p (noting that $U.$ and $C_p(X_N)$ have to be recomputed at each step), until the value of C_p achieves a satisfactory target magnitude.

DISCUSSION

The subset of examinees (or items) extractable by the technique described in the preceding section -- or, indeed, by all earlier methods, to our knowledge -- is unidimensional (or nearly so) in the scalability or order-theoretic sense. On the other hand, the unidimensionality of data required for satisfactory practical functioning of, and parameter estimation in, ICC models is more closely related to that in the factor-analytic sense.* Since it is well known (e.g., Guttman, 1950; DuBois, 1970) that scalability of a set of items leads to a simplex and -- depending on the distribution of difficulties of the items -- may produce a correlation matrix with up to n/2 factors, it may seem meaningless to strive for unidimensionality in the sen'e of scalability when the purpose is to improve the applicability of ICC models.

Despite the foregoing circumstance, experience has shown that when a set of items approximates scalability in a given group of examinees, the factorial structure also becomes "cleaner" and the estimability of the latent trait parameter is improved. This is described in detail by Birenbaum and Tatsuoka (1980), who found that all these improvements -- scalability, factorial determinancy and estimability of θ -- result simultaneously when item scores are

---

*Lord and Novick (1968, pp. 374-375), show that the matrix of item tetrachonics having unit rank (when communalities are inserted in the diagonal) is a <u>sufficient</u> but "very far from being (a) necessary" condition for the assumption of local independence of the items to hold.

modified by assigning zeros to those items that were deemed (by an elaborate system of error analsyis) to be correctly answered for wrong reasons. Thus, improving unidimensionality in the scalability sense -- or, to put it another way, removing examinees with aberrant response patterns -- does enhance the practical applicability of ICC models up to a certain point. But there are limits to the efficacy of this approach, which are discussed elsewhere (Tatsuoka & Tatsuoka, 1980).

We now turn our attention to a couple of difficulties with the NCI that we have yet to resolve to our satisfaction. The first is the excessively small (i.e., close to -1) value received by a student whose test score would have been perfect except for his/her getting one or two very <u>easy</u> items wrong by mistyping or some other clerical error. For instance, consider the response pattern (111111111101), whose NCI value is -.818. Yet, this students getting the second easiest item wrong is almost certainly due to a random clerical error, and hence the response pattern should not be regarded as "extremely atypical" in the sense of its implying a serious misconception. In particular, it seems incongruous that such a response pattern should be automatically deleted from the outset in the method for extraction of unidimensional subsets discussed in the preceding section. Thus, this extreme sensitivity of the NCI to one or two "happenstantial" zeros in a response pattern that would otherwise receive a value of +1.0 is an undesirable property so long as we adopt overall group consistency as the criterion for extracting unidimensional subsets.

Fortunately, however, the above defect of the NCI does not affect its usefulness in the diagnostic procedure utilizing the fourfold table, based on the NCI, ICI and score combination, displayed in Table 4. It will be recalled that, so long as the total score is high, a student with low NCI will not be diagnosed as having serious problems even when the ICI is high. It is also seen from Table 4 that, before any diagnosis can be made on the basis of the NCI's being high or low, it is essential to examine whether the total test score is high or low.

All in all, it appears that the ICI is the more useful of the two indices for diagnostic purposes. Its drawback is that it requires the test to be constructed out of two or more parallel subtests. Alternatively, we might say that, for achievement tests to perform as powerful diagnostic tools, they should incorporate several parallel subparts.

It was pointed out earlier that Equation (2) fails when the test score is either zero or perfect, making $U = 0$. Intuitively, C_p should have the value 1 in both these cases, since a response vector with all elements equal 0 and one with all elements equal 1 are both Guttman vectors. This can be brought about in one of two ways: (a) by arbitrarily defining $C_p = 1$ when $U = 0$ despite the fact that Equation (2) does not apply in this case -- much as we define $0! = 1$ even though the definition $n! = 1 \cdot 2 \cdot 3 \cdots n$ does not make sense for $n=0$; or (b) by changing the very definition of C_p to read

(5) $$C_p = \frac{2(U_a+1)}{U+1} - 1$$

instead of $C_p = 2U_a/U - 1$ as in Equation (2).

Alternative (a) has the advantage of preserving the definition of Cliff's consistency index C_{t1} as a weighted average

$$\sum_{j=1}^{N} w_j \cdot NCI_j$$

with

$$w_j = U_j / \sum_{i=1}^{N} Ui$$

as stated in Equation (3). For any pattern with $U_j = 0$, w_j would be 0 and hence the value $NCI_j = 1$ would not enter into the weighted average for computing C_{t1}. This is consistent with Cliff's definition, because perfect and all-zero response patterns do not contribute anything to a dominance matrix, since no item dominates any other item in such response patterns.

On the other hand, alternative (a) has the disadvantage of rendering undefined the ICI for a student with perfect (all all-zero) response patterns on all of the parallel subtests, for the combining weights for all the (individually perfect) C_p's would then take the indeterminate form 0/0. Thus, it would require another definition by fiat to give such a student's ICI the value of 1, which is what is should be, since all the response patterns are identical.

By contrast, alternative (b) would avoid this difficulty. With the revised definition (5) for C_p, the combining weight associated with the C_p of the j-th among a set of m response patterns would be

$$(6) \qquad w_j = \frac{U_j+1}{\sum_{j=1}^{m} U + m} \quad .$$

Consequently, the combining weights used in calculating the ICI for a student who consistently shows perfect (or all-zero) response patterns on m parallel subsets would uniformly equal $w_j = 1/m$. Hence his/her ICI will now be 1.0, as it should be. On the other hand, alternative (a) would lead to an overall group consistency index that does not agree with Cliff's C_{t1}, since each NCI would no longer be a linear transform of Ua/U.

Thus, each of the alternatives for making NCI take the value 1.0 for perfect and all-zero response patterns has its advantage and its disadvantage, and we have a dilemma. In view of the more important role played by the ICI compared to the NCI in error diagnosis, we are inclined to favor alternative (b). However, further investigation of other possible implications carried by definition (5) for pattern conformity needs to be made before we make a definite commitment.

REFERENCES

Birenbaum M. & Tatsuoka, K.K. The use of information from wrong response patterns (Technical Report 80-1). Urbana, Ill.: University of Illinois, Computer-based Education Research Lab, 1980.

Brown, J.S., & Burton, R.R. Dignostic models for procedural bugs in basic mathematics skills. Cognitive Science, 1978, 2, 155-192

Cliff, N. A theory of consistency of ordering generalizable to tailored testing. Psychometrika, 1977, 42, 375-399.

Davis, R.B. The Madison Project discovery in mathematics: A text for teachers. Reading: Addison-Wesley, 1964.

DuBois, P.H. Varieties of psychological test homogeneity. American Psychologist, 1970, 25, 532-536.

Guttman, L.L. Studies in social psychology in World War II. In S.A. Stouffer (Ed.), Measurement and Prediction, Vol. IV. Princeton: Princeton University Press, 1950.

Krus, D.J. Order analysis of binary data matrices. Los Angeles: Theta Press, 1977.

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

Mokken, R.J. A theory and procedure of scale analysis: With applications in political research. The Hague: Mouton, 1971.

Reynolds, T.J. The analysis of dominance matrices: Extraction of unidimensional contest (Technical Report No. 3). Los Angeles: University of Southern California, Department of Psychology, June 1976.

Tatsuoka, K.K., & Birenbaum, M. The danger of relying solely on diagnostic testing when prior and subsequent instructional methods are different (Technical Report 79-2). Urbana, Ill.: University of Illinois, Computer-based Education Research Laborotory, 1979.

Tatsuoka, K.K. The least-squares estimation of latent trait variables. Paper presented at the 1979 American Educational Research Association, Boston.

Tatsuoka, K.K., & Tatsuoka, M.M. Consistent errors, dimensionality and local independence (Technical Report 80-6) to be prepared.

Tatuuoka, K.K., Birenbaum, M., Tatsuoka, M.M., & Baillie, R. A psychometric approach to error analysis on response patterns (Technical Report 80-3). Urbana, Ill.: University of Illinois, Computer-based Education Research Lab, 1980.

Van Der Flier, H. Environmental factors and deviant response patterns. In Poortinga (Ed.), Basic problems in cross-cultural psychology. Amsterdam: 1977.

Wood, R.L., Wingersky, M.S., & Lord, F.M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters (Research Memorandum 76-6). Princeton, N.J.: Princeton University, Educational Testing Service, 1976 (revised 1978).

Yamamoto, Y. & Wise, S. Extracting unidimensional chains from multi-dimensional datasets: A graph theory approach (Technical Report 80-2). Urbana, Ill.: University of Illinois, Computer-based Education Research Lab, 1980.