Clustering Binary Items

Douglas McCormick
Norman Cliff
Robert Cudeck
Thomas Reynolds

Department of Psychology
University of Southern California
Los Angeles, California 90007

Technical Report 81-1
March, 1981

Funding for this research was provided by National Institute of Justice Grant # 79-NI-AX-0065.

The authors very gratefully acknowledge the assistance of Judith Toben Zatkin and Linda Collins in preparation and discussion of this manuscript.

Abstract

The overall goal of this paper is to describe a new method of finding patterns in binary data, and to report the results of a number of experiments designed to evaluate the procedure with artificial data of a realistic type. This report is one of several in a series focussing on this problem or applying the methods to several important sets of data.

DIFFICULTIES IN CLUSTERING BINARY DATA

The division of a set of binary measures into roughly unidimensional subscales has a close resemblance in purpose to the subdivision of a set of continuous measures through factor analysis. Unfortunately those methods which serve so well when applied to variables which take on many values are rife with snares and pitfalls when our goal is to divide binary measures into homogeneous subscales.

First among the dangers inherent in a conventional factoring have to do with the indices which measure association among the variables. Correlations, covariances and cross-products may allow sensible analysis of continuous measures for a variety of purposes, but their possible application to binary measures is troubling because they reflect not only agreement in the basic measurement of two variables but also the similarity in the frequency of 1's. For example, suppose one has identical continuous number series which are dichtomized differently for each series. One may obtain correlations which range from unity almost to zero simply by changing the value for each series below which all numbers become zeros and above which they become ones. Fairly modest differences in the two frequencies of 1's in the two measures are known to result in substantial decreases in their correlation.

The second difficulty, if one intends to factor analyze

binary data, is that if one abandons the questionable product moment indices, the results of factoring less conventional indices can no longer be interpreted as representing variance accounted for. Since the whole rationale for factor analysis is to account for the maximum possible variance with a reduced number of variables, the entire process becomes problematic. This can be especially true if the results are presented to an audience which attributes to them a conventional factor analytic interpretation.

In seeking alternatives to conventional factor analysis Christofferson (1975) has turned to a hypothesis testing method developed by Bock and Lieberman (1970). However, in addition to severe practical limitations in the number of items which may be analyzed, it offers no alternative to the researcher who wishes to discover the structure of his data rather than test a particular hypothesis. Hofmann (1980) has taken the route, just criticized, of applying a non-conventional index to a conventional analysis and he offers similar cautions concerning interpretation of results.

An alternative to these approaches, which will be described here, is to select an index suitable for binary measures and to increase the homogeneity of subsets among these measures through an agglomerative process which clusters together variables measuring similar properties. This type of procedure is in a class which has been collectively described as cluster analysis methods. Although cluster analysis of binary measures is quite common, the data are usually taxonomic

criteria such as "has feathers vs. has no feathers" which do not reflect measurement of an underlying continuum. In fact the usual purpose of taxonomic cluster analysis is not to find groups of variables which go together but to create a tree structure which organizes various life forms measured by the variables. Application of cluster analysis to the problem of isolating homogeneous subscales of binary items appears to be a novel idea.

Cluster Analysis for Binary Items

There is enormous literature on cluster analysis methods. Blashfield (1980) identifies three quite distinct subliteratures within psychology on cluster analysis, one stemming from Tryon, one from Johnson, and one from Ward, although the Tryon tradition (Tryon and Bailey, 1970, e.g.) is actually more like a form of factor analysis. In addition, there is a virtually separate, biologically oriented field, exemplified by the books of Sokal and Sneath (1963; Sneath and Sokal, 1973). Within it, there is also a substantial amount of differentiation and compartmentalization. Fortunately, there are available several books that go some distance toward providing an organization and unification of the literature (Hardigan, 1975; Everitt, 1977).

At a broad, conceptual level, the various clustering methods may be viewed as attempting to group entities of some sort in such a way that there is considerable homogeneity within groups and heterogeneity between them. Within that general scheme, every clustering procedure consists of the same set of

conceptual parts, but each of the parts takes a different form according to the particular method. Every clustering method uses an index of homogeneity or similarity, whether it takes the form of a Euclidean distance, a Pearson correlation, or some more exotic index. Every method makes a structural assumption. This is an assumption about the mode of organization of the entities being clustered. In taxonomy, the organization is heirarchical because it is assumed that this is the correct structure for taxonomy. In others it is assumed the structure is disjoint but not necessarily hierarchical, and in others the clusters are allowed to overlap. This assumption is quite fundamental to the process, and there is probably nothing more damaging to a cluster analysis than to impose an inappropriate structural assumption.

Finally, each clustering procedure has an algorithm, a procedure by which clusters are constructed. The nature of this procedure depends to some extent on the homogeneity index and the structural assumption, but there are some common strategies and some commonalities across methods. The method may start with individual elements and work by accretion or it may start with the whole mass and subdivide it. It may base its decision on how to add subsequent entities by finding the single most similar or dissimilar one, or may use some form of average of a number of them. That is, it may use single linkage (minimum diameter), complete linkage (maximum diameter) or average linkage. There are also a large number of details within each algorithm which may differ from method to method. The choice concerning the algorithm to use is based partly upon the structural assumption and partly on considerations of

computational efficiency.

From all this it is reasonable to conclude that a clustering procedure should be tailored to the special characteristics of the data on which it is employed. The index of homogeneity that is used should reflect the special nature of homogeneity that applies in a given context. The structural assumption should be carefully chosen as valid for the entities being clustered. The algorithm should be consistent with the structural assumption and should combine efficiency with robustness in the face of error.

This research was undertaken with the idea that available clustering methods were not well suited to the clustering of the data that were of primary interest to us. These data consist of entities (persons) by variables matrix, where the scores are binary (dichotomous) and the object is to group together variables into homogeneous subsets. Subsequently, the persons may be given scores on the scales defined by the subsets of variables or items, but this is a secondary step. The subsets need not be disjoint, although this is desirable, and they certainly need not be hierarchical, although this would be interesting. Thus, the overall goal we have in mind is more like that of factor analysis than like taxonomy.

The program whose workings are described here is tailored for use with this kind of data. Considerable thought was given to the form of the index of association to use, and the program includes several optional indices. The clustering strategy used is a particular form of agglomeration which is based on a kind of

average linkage; it imposes no restrictions concerning disjointness of the final clusters. The specifics of these features are described in the pages that follow, along with an extensive Monte Carlo study designed to test the overall efficacy of the method and the relative success of different indices and options.

Development of a clustering method to handle binary data came about in several steps. Refinement of the clustering algorithm allowed a closer examination of indices which led to more analyses, which in turn led to further refinement of the algorithm and so on. Once the method for clustering reached a sufficient level of refinement, the present evaluation of different indices seemed warranted using a systematic collection of artificial data sets.

Indices of Association

Traditional Indices

Of the multitude of indices of association which may be applied to data composed solely of ones and zeros, most are felt to have defects for defining subscales or factors. Most, including the product moment indices and the χ^2 association measure, were initially devised for very different purposes, and their application in this context, though mathematically possible, has remained conceptually clouded at best (Carroll, 1961).

There are three common measures of association that can be considered as the bases for defining scales with binary data; two that measure association between pairs of variables, and one measuring the overall consistency among a groups of variables.

The product-moment correlation, which is the phi coefficient in the binary case, may be computed between pairs of variables, and the consistency of a set can be measured by the average inter-item correlation. It has the apparent disadvantage that the phi coefficient is influenced by the similarity of frequencies of the two items; thus, it is likely that variables of equal frequency tend to cluster together, over and above the influence of similarity of what they measure. Nonetheless, the Pearson r is so familiar and so widely used, that it is essential to include it as one alternative for forming clusters.

An index that is not influenced by similarity of marginal frequencies is the Goodman-Kruskal (Goodman and Kruskal, 1954) gamma (γ). If Table 1 is used to refer to the joint frequencies of two variables, the γ coefficient is defined as

$$\gamma = \frac{ad - bc}{ad + bc} \quad . \tag{1}$$

It can have the range ± 1.0 regardless of the item marginals, and is zero when the items are independent. It has two other names in the literature; one is Yule's q (Yule, 1912). It is also a specialized form of Kendall's tau with ties (Kendall, 1970). The average inter-item gamma is a promising index on which to form clusters.

The third index which suggests itself as a measure of association among a group of items is the Kuder-Richardson Formula 20 coefficient of internal consistency (e.g., Guilford and Fruchter, 1978; Lord and Novick, 1968; Nunnally, 1970). It is defined by the formula

Table 1

Schematic Fourfold Table

		Item 2: Correct	Item 2: Incorrect	
Item 1	Correct	a	b	n_1
	Incorrect	c	d	$n-n_1$
		n_2	$n-n_2$	

$$KR20 = \frac{k}{k-1} \frac{S_x^2 - \sum_{j=1}^{k} S_j^2}{S_x^2} \quad , \qquad (2)$$

where S_x^2 is the variance of total scores across the k items in the cluster and S_j^2 is the variance of an individual item in it, i.e., P_j (1 - P_j) where P_j is the proportion endorsing or passing item j. Clustering items so as to maximize the within cluster KR20 is also a plausible strategy, and so it too was studied. Like the Pearson r or ϕ coefficient, similarity of item marginal frequencies influences KR20, so it too may have a tendency to form clusters on this basis. It also has the disadvantage that it tends to increase with the number of items in the cluster, so it is difficult to ascertain cluster boundaries when it is used.

These three indices are implemented in the clustering program studied here, along with a fourth to be described in the next section. There are a number of others that might have been used, but we feel that these are the ones that offered the greatest promise. A notable omission, perhaps, is Guttman's coefficient of reproducibility (Rep). It is omitted because of the controversy that has surrounded it since its early days (Green, 1954); White and Saltz, 1957) and also because of the complexity of programming it and the computer time involved. The clustering strategy, however, could be applied with Rep or any other form of association index.

Quality Indices

Recently Cliff (1979) has provided a fresh

conceptualization of binary indices which allows many to be described in a common framework based on order relations. Cliff also suggests that a heretofore untried subclass of indices may have properties useful for selecting binary variables (or items) to make a unidimensional scale.

The Conceptual Framework

We assume that information on a group of persons or entities is available for a sample of variables. The basic unit of information then pertains to a rating of the person or the variable according to any of a variety of plausible experimental procedures. For example, a particular datum may arise by rating a person on a variable, with a "1" denoting that a certain trait applies to the person, and zero denoting the converse. In ability testing, a person who correctly answers a question scores 1, and receives a zero otherwise. In criminology, we record the fact that a particular juvenile has engaged in a specific kind of crime by a 1, and use zero to show that no such behavior has been noted. It is clear that each of these experimental procedures shares the common feature of providing the most basic form of recording the presence or absence of traits for the persons.

In his approach to association indices, Cliff begins with the order relations created as each variable divides people or objects into two categories. Each member of the group scored 1, is ordered ahead of each member of the group scored 0, but within each of the two groups no order is established.

This is the description of a single item's order relations. When a second item has created its ordered groups, the order relations existing for any two persons must fall into a limited set of arrangements according to their joint relations (see Table 1). If the pair of persons is not ordered by either item, no relations are available at all. If the pair is ordered identically by both items (one person is scored 1 on both, while the other is twice scored 0) the orders are said to be "Redundant." If one item distinguishes between the persons while the other does not, the ordering item is said to provide a "Unique" relation relative to the non-ordering item. If the ordering of one item is opposite that of the second item then they are "Contradictory."

Between all possible pairs of items, these three kinds of order information--redundant, unique and contradictory--can be calculated. Now one purpose to which this information can be applied is to identify subsets of variables among binary data. For example, it is of interest to know to what extent patterns of offenses may occur in criminal records. Surely what one would seek is a categorization of offenses which are internally consistent and which provide differentiation among the offenders. In the present terminology, such a categorization would be characterized by having a small number of contradictory relations relative to the redundant ones, and by having enough unique relations to provide differentiation. To produce such a categorization, an index is sought which utilizes all three sources of information.

The class of indices we will recommend below require that the redundant, unique and contradictory relations be aggregated formally. One strategy (which we do not recommend!) is to simply count the number of relations of the three types which occur in some data set. Considering all the possible combinations of items for a large sample of 1000 persons and 100 offenses seems an impossible task--nearly 2.5 billion comparisons! Fortunately, it turns out not to be necessary to follow this procedure. Instead, the required information can be deduced from summary information such as is contained in Table 1.

The product ad in Table 1 is the number of relations that are redundant between the two items, and bc is the number that are contradictory between the two. The sum ac + bd is the number unique to item 1, whereas ab + cd is the number unique to item 2. Thus the basic quantities are derivable from this information. We let $\underline{r}_{jk}$, $\underline{c}_{jk}$, and $\underline{u}_{jk}$ stand for the number of redundant, contradictory, and unique relations on a given pair of items. Then, considering all pairs of items within a particular set

$$r = \sum_j \sum_k r_{jk} \quad ; \tag{3}$$

$$c = \sum_j \sum_k c_{jk} \quad ; \tag{4}$$

$$u = \sum_j \sum_k u_{jk} \quad . \tag{5}$$

Using only these quantities, it is possible to express Pearson's r, the Goodman-Kruskal Gamma, and KR20;

$$\text{Pearson } r = \frac{r_{jk} - c_{jk}}{\sqrt{(r_{jk} + c_{jk} + u_{jk})\ (r_{jk} + c_{jk} + u_{kj})}} \qquad (6)$$

$$\text{Goodman-Kruskal Gamma} = \frac{r_{jk} - c_{jk}}{r_{jk} + c_{jk}} \qquad (7)$$

$$\text{KR20} = \frac{x(\Sigma r_{jk} - \Sigma c_{jk})}{x\Sigma r_{jk} - (x - 2)\Sigma c_{jk} + \Sigma(u_{jk} + u_{kj})} \qquad (8)$$

where x = the number of crimes or items.

One may note that similarities exist among these three indices regarding the treatment of the three types of relations. The numerator in each case is either the difference between Redundant and Contradictory relations or, in the case of KR20, a multiple of this figure. The denominators of the three indices are generally more complex, but a major difference between them is the inclusion of Unique relations in the denominator of the two product moment indices, KR20 and r, but not in the denominator of the Goodman-Kruskal Gamma. The presence of Unique relations in the denominator accounts for the well-known property of r to be reduced when the frequencies of the variables are different, and the equally well-known fact that KR20 reaches its maximum value only when all items in a set have equal frequencies. In the psychometric literature this topic has been examined under the name of the attenuation paradox. This property of the product moment indices makes ordinary factor analysis suspect when applied to binary measures, since the association due to common levels of item frequency act to confound the association due to measurement of common properties. When the effects of

item frequency entirely overwhelm those of substantive measurement, the resulting factors in psychometrics are spoken of as "difficulty factors" (Carroll, 1961), and they are of course useless for any of the purposes to which factors are ordinarily applied.

By excluding Unique relations in its denominator, Gamma handles all items without regard for their frequencies. Although this would seem to be far better than the approach offered by r and KR20, there is another possibility which can be explored.

One of the problems with describing associations between items with the product moment indices is that, in the extreme case, items of a uniform frequency tend to separate all subjects into two homogeneous groups. Items such as these do not provide sufficient differentiation among the subjects, and are usually undesirable for that reason. Gamma, although it does not promote the accumulation of items at the same frequency, does nothing to prevent it, or to recover a cluster with items spread out at a variety of levels. Within the framework discussed here this can be accomplished by putting Unique relations, which encourage clustering items with a broad range of frequencies, into the numerator of a new index and to give them a positive weight. One such index is shown in Equations (9) and (10). In Equation (9) t is the familiar difference between Redundant and Contradictory relations

$$t = \Sigma r_{jk} - \Sigma c_{jk} + .25\Sigma u_{jk}$$

$$t_{jk} = r_{jk} - c_{jk} + .25u_{jk} + .25u_{kj} \qquad (9)$$

but now with the addition of the Unique relations with a modest positive weight of .25. The quantities in Equation (9) are defined for a single pair of items. They may be summed across all the pairs of items in a scale to form an overall total t. The possible combinations of weights for all three relations are infinite and this combination is necessarily arbitrary, but it does retain the character of Gamma with the addition of a fraction of the Unique relations to test their effect. t_b and t_c in Equation (10) serve to scale the index q between a maximum of 1 and a minimum of minus 1 with a rational zero point which represents statistical independence. t_b is the best or maximum t which would occur in a Guttman scale having the same marginals as the actual data. t_c is the value of t which would be expected if the items were unrelated.

The three indices, q, gamma and r are included here because they represent different levels of encouragement of measurement at different levels of difficulty: a positive attitude (q), a neutral attitude (gamma), and discouragement of measurement at different levels of difficulty (r). KR20 is also included although it behaves similarly to r, because of its widespread influence on the evaluation of binary items in the testing field.

$$q = \frac{t - t_c}{t_b - t_c} \qquad (10)$$

The Clustering Algorithm

The clustering process was designed to occur in an agglomerative fashion (Sneath and Sokal, 1973). Each of the variables begins its own cluster, and individual variables are added to each of the original variables until each cluster contains the entire set of variables. At any point in the accretion process the variable added is the one which has the highest average index with each of the variables already contained in the cluster. In the selection of the best variable to add, no attention is paid to whether that variable is included in any other cluster. With well-defined clusters, it can be expected that clusters will duplicate each other at some point. For example, a cluster consisting of variables 1 and 2 adds number 3, and one consisting of 2 and 3 adds item 1. They are the same, and their subsequent histories must be also. This is important for judging the clarity of a solution. Each of the cluster histories is formed independently from all the others. Each history is described by a set of variable names or numbers and the average index value when it entered the cluster.

Because this process results in completely overlapping sets of variables, it cannot be represented by the familiar tree-like dendrogram of taxonomic cluster analysis. Unlike the biologists who wish to find a tree-like structure in the data, the purpose here is to define subscales, so the loss of the tree diagram should not trouble us. What is required however, is a "stopping rule" for deciding when the end of a cluster has been reached and later items do not belong, short of the

point where all the clusters contain all the items.

Currently, we use two methods for stopping the clustering process. One method uses the often observed fact that if clusters exist in the data which are in an intuitive sense genuine, then in spatial terms, areas of high density will be separated by "moats" of empty space. When all the variables within a region of high density belonging to a cluster have been included, reaching across the moat for the next best variable should be accompanied by a large drop in the average index for this entering variable. One method, then, of stopping the cluster history at the proper point is to cut the list of variables where the largest drop in the average index occurs (Wainer and Schacht, 1979).

When the data do not exhibit such clean-cut clusters, but rather consist of only vaguely defined areas of high density separated by regions of gradual thinning and rethickening, a more arbitrary approach must be adopted. When this is the case, an absolute value is selected for the average index, below which no item may be added.

An Example

Table 2 illustrates the clustering process using a small set of computer-generated data. The data consists of eight hypothetical items. The first six belong in a cluster and the last two are expected to be outliers. The first 8 x 8 matrix shows the q indices between the items. In the second matrix is the record of average q values as each successive item is

Table 2

An Artificial Example with Eight Variables

1.00	.22	.14	.31	.17	.24	-.03	.01	
.22	1.00	.22	.13	.16	.13	-.05	.06	
.14	.22	1.00	.22	.17	.22	-.02	-.07	
.31	.13	.22	1.00	.09	.17	.01	-.09	
.17	.16	.17	.09	1.00	.13	.00	-.12	Q
.24	.13	.22	.17	.13	1.00	-.26	-.14	
-.03	-.05	-.02	.01	.00	-.26	1.00	-.05	
.01	.06	-.07	-.09	-.12	-.14	-.05	1.00	
1.00	2.00	3.00	4.00	5.00	6.00	7.00	8.00	
.31	.22	.22	.31	.17	.24	.01	.06	
.21	.23	.21	.21	.20	.25	.14	.09	
.19	.19	.21	.19	.18	.19	.12	.12	Record of Q's
.17	.19	.17	.17	.18	.17	.14	.12	
.15	.15	.15	.15	.17	.15	.12	.13	
-.05	-.05	-.05	-.05	-.05	-.05	.12	.10	
-.06	-.06	.06	.06	-.06	-.06	-.06	-.05	
1	2	3	4	5	6	7	8	
4	1	4	1	1	1	4	2	
6	4	1	6	4	4	1	1	
3	3	6	3	6	3	3	4	
2	6	2	2	3	2	2	3	Item History
5	5	5	5	2	5	5	6	
7	7	7	7	7	7	6	5	
8	8	8	8	8	8	8	7	
1	1	1	1	1	1	1	1	
1	1	1	1	1	1	1	1	
1	1	1	1	1	1	1	1	
1	1	1	1	1	1	1	1	Membership by
1	1	1	1	1	1	1	1	Largest Cap
1	1	1	1	1	1	1	1	
0	0	0	0	0	0	1	0	
0	0	0	0	0	0	0	1	
1	1	1	1	1	1	0	0	
1	1	1	1	1	1	0	0	
1	1	1	1	1	1	0	0	
1	1	1	1	1	1	0	0	
1	1	1	1	1	1	0	0	Membership by > .10
1	1	1	1	1	1	0	0	
0	0	0	0	0	0	1	0	
0	0	0	0	0	0	0	1	

added to a cluster. Each column represents the cluster begun by the item whose number is recorded in the first row. Below the initial item number in each column is the average q value for the next item added, then the third item, then the fourth and so on until all eight items are recorded for all eight clusters.

In general the clustering process proceeds from higher to lower index values as it does for the cluster recorded in column one of the second matrix. However, if a cluster begins at a point of relative low density and moves into a region of high density, the index may reverse direction and rise. This condition is most clearly demonstrated in the record of clusters seven and eight. Because these items are essentially unrelated to the others, the first additions to their clusters are made with very low indices, .01 and .06. Subsequent items obtain higher averages primarily due to their association with each other and in spite of near zero relations with items seven and eight. The pattern demonstrated by items seven and eight then is an indication, when it occurs, that the item beginning the cluster is an outlier.

The third matrix is a record of the clustering which indicates the identifying number of the items added with the index values shown above in the second matrix. Here it is clearly shown that items 7 and 8 are the last to enter every cluster except the ones which each began.

The final pair of matrices which contain only ones and zeros represents the items retained in each of the eight clusters after selection of a cutoff or stopping point, beyond

which no new items were accepted. The columns in these matrices refer, as in matrices 2 and 3, to clusters begun with items 1 through 8. The rows this time refer to items 1 through 8 also so a one in row 7, column 7, means item 7 was included in cluster seven before the stopping rule intervened.

Matrix 4 is the membership which resulted from stopping each cluster where the largest drop in the average index occurred. All the clusters contain items 1 through 6 and cluster seven also contains item 7. Cluster eight also contains item 8. One might conclude at this point that a legitimate cluster consists of items 1 through 6 and that 7 and 8 entered clusters only because they were the starting items. The low index values with which clusters seven and eight began should also have indicated to the investigator that they represented outliers. Nevertheless, matrix 5 displays the correct cluster solution without requiring any such reexamination. Matrix 5 is the result of simply cutting off the clusters when the indices dipped below .10. Here items 7 and 8 are clearly outliers and 1 through 6 form a solid block as before.

Evaluation with Artificial Data

Evaluating the success of a clustering scheme is generally aided by use of artificial data analyses. Because the data can be generated from a known underlying cluster structure, comparison of the cluster results to the known underlying model may be relatively unambiguous compared with an analysis of empirical data where a plausible structure must be assumed

according to the subjective judgment of the researcher. Even with artificial data, the evaluation remains somewhat ambiguous for the majority of clustering methods that do not provide an objective stopping rule. When such methods are evaluated, typically the researcher allows himself the privilege of selecting the cluster solution which best fits the correct answer. Since this benefit is allowed for all competing methods no predictable bias is present. Such a procedure neglects the real question, however, of how well any such methods will behave when the optimal stopping point must be deduced from the data alone.

Simulation Model

In order to examine these methods with artificial data, one of course needs a model by which to generate simulated persons and variables. An appropriate model would be one which plausably represents many of the phenomena to which the method might apply. As mentioned earlier, such data may arise from rating the presence or absence of certain traits in a sample of individuals, or may occur when a group of subjects responds to items in a questionnaire, or may be represented in the offense histories of a sample of delinquents. The common features of such data are that the observed binary variables should be based on known underlying clusters, that there be a mechanism for specifying very clear clusters, very loose ones and essentially independent variables, and that there be a process for injecting a random component of error into the data. A number of latent trait models exist in the social sciences which fit

these requirements, but one with which we have had experience is the Birnbaum (1968) latent trait model for psychological tests.

This model postulates that a binary variable is based upon an underlying probabilistic function of four components, only three of which are relevant here. It states

$$p_{ij} = \frac{1}{1+e^{-1.7a_j(g_i-b_j)}}$$

with

$$x_{ij} = 1 \text{ if } p_{ij} \geq t$$

$$x_{ij} = 0 \text{ if } p_{ij} < t,$$

where

p_{ij} = the probability, ranging from 0 to 1, that person i will receive a score of 1 on variable j

g_i = the ability of person i

b_j = the difficulty of item j

a_j = a consistency or discrimination score for variable j, which pertains to the precision of its cluster

x_{ij} = the observed score for person i on variable j, which is set to 1 if p_{ij} is greater than a stochastic threshold t.

t refers to the myriad external things that can influence whether a high p_{ij} is actually recorded as 1. For example,

a rater may be temporarily distracted when a behavior should have been recorded, as a juvenile about to snatch a purse might be deterred by viewing a passing patrolman. The person's ability, g_i refer to the overall likelihood that a given individual will tend to be rated 1 on all variables. For example, highly gregarious types are extremely likely to be rated 1 on all variables of a sociability questionnaire, whereas "wallflower" types are unlikely to receive many ratings of 1. In criminology, recidivists with histories of many felonies have a general propensity for crime, whereas for a group of social workers even mild misdemeanors may be infrequent.

The difficulty (or easiness) scores for variables, b_j, are exactly like the ability score for persons. They index the overall probability that a variable will be scored 1 in some population. For example, in a certain population, some variables from a sociability questionaire may almost always be checked for the persons, whereas other variables might be checked infrequently. Criminologically, one might suppose that driving while legally intoxified occurs relatively frequently, while other offenses, such as homicide, have a characteristic infrequency.

It is clear that the parameters g_i and b_j must be defined relative to each other. A given set of g_i will be judged as extreme as the distribution of b_j is shifted. For instance, the frequency with which a given sociability variable is checked depends upon the sample of persons who are rated. If the sample contains mostly extroverts, a variable's frequency will be much

higher than when the sample contains mostly introverts. Similarly, if a juvenile is rated on a self-report questionnaire of delinquency, he may show a high propensity on status offenses, but a near-zero propensity on index charges.

In the simulation model, the scores for g_i and b_j may range over the domain of real numbers, and may assume any convenient mathematical distribution. Since g_i and b_j are scaled relative to each other, in the following studies we arbitrarily fixed the g_i to have a mean of zero and standard deviation of 1, and modified the b_j in relation to this standard. For the most part we assumed that the distributions were normally distributed, but at times produced scores for the latent g_i and b_j which were uniformly distributed or bimodal. It is possible to also sample a_j, the index of consistency of the variable within its cluster, from a variety of distributions, but for the most part we fixed these parameters to values of .5, 1 and 3, corresponding to low, medium and highly consistent clusters, respectively.

Taken as a whole, it can be seen that this simulation model is adequate for our purposes. It allows a variety of different kinds of data sets to be generated which have properties much like the real data which clustering methods might analyze. We also think that these properties of the model are sufficiently general as to be common in many fields of the social sciences. Manipulating the means, standard deviations, and distributions of the parameters in the model allows the simulation of an infinite variety of types of data. In our

experiments a few of these possibilities were selected as interesting or realistic cases, and thus no attempt was made to cover an exhaustive set of possible parameters.

Procedure

In the experiments which follow, a common procedure was used. We generated data which produced 24 variables, assigned as follows: variables 1 through 4 came from one cluster, variables 5 through 10 formed a second cluster, variables 11 through 18 made up a third, and variables 19 through 24 were completely independent of the three clusters and of each other. With the true membership of these variables always known, we could judge the accuracy of the method by noting the number of items correctly placed in their original clusters.

A particular data set was created in steps. First we generated a distribution of g_i, for each cluster independent, most frequently using a sample size of 500 cases. Next we decided upon typical values for a_j. These values were fixed for all variables. Then we sampled a group of 4 b_j, then independently sampled a second group of 6, then produced another independent group of 8, then six additional independent values for the singletons in the data set. With these vectors of initial parameters, the Birnbaum model produces a data set of 500 rows and 24 columns of p_{ij}. Then using a process for uniform random numbers in the interval (0, 1) denoted t, we created the matrix of binary values by scoring 1 if $p_{ij} \geq t$ and zero otherwise. This matrix of binary

values was used as the raw data in the clustering program. For every condition in the experiments we generated several independent data sets using the above procedure and recorded the program's performance over all replications.

Dependent Variables

Two types of dependent variables were devised for this study. The present method is a non-hierarchical cluster analysis in which each variable begins a cluster, and other variables are added to it according to the highest within-cluster average. Typically, the average within-cluster index value decreases as items of decreasing similarity are added. At some point an optimal or statistical decision must be made regarding where the true cluster members have stopped entering, and outside members have begun to intrude. One method, the optimal one, is to stop all clusters known to have k legimate members when they contain k members of whatever identity. This rule is applied to all clusters, and in the simulation case, one may then ascertain how many variables have been correctly recovered in each cluster. Typically, in any single cluster the members thus identified will contain a preponderance of correct variables, and perhaps a few false choices also. This first method uses only the number of correctly selected variables, expressed as a proportion of the true number known to be in the cluster.

A second dependent variable used the largest gap rule discussed above. Here we employed a very stringent rule which required that the largest gap exactly identify only the true

cluster members and never include a non-member. Considering the fact that this rule required the true cluster to be identified only statistically it can be appreciated that this procedure is very stringent indeed.

Computer Program: BINCLUS

We operationalized the clustering procedure into a computer routine called BINCLUS (for Binary Clustering). BINCLUS consists of three distinct phases. The first part generates the original raw data of binary variables according to the parameters described above, and calculates the redundant, unique and contradictory order relations among all items. The second step uses these r's, u's and c's to construct one of four kinds of association index; Pearson r (which is also the Phi coefficient with binary data), the Goodman-Kruskal gamma, the Kuder-Richardson 20 coefficient, and the quality index, q. The third section which is the heart of the program, performs the non-hierarchical clustering. The clustering routine can be applied to any convenient set of association indices, so each of the four measures were used in some of the simulations.

The following sections describe three experiments. These represent a logical progression, with each experiment elaborating and extending the findings of the former studies. In certain studies our interest centered around the performance of the clustering algorithm, and addressed the question: Does the method correctly recover clusters of known structure? At other times, our interest was in the relative performance of q

compared to the other indices. At these times the question became: What is the relative performance of the q index given its use in our non-hierarchical clustering method?

Experiment 1

Purpose of the Study

The first study was intended to test the effectiveness of BINCLUS using only the q index. It will be remembered that q includes not only positive weight for redundant and negative weight for contradictory order relations, but also a positive weight for unique relations. Its use, then, is designed to cluster groups of variables insofar as they approach the ideal of the Guttman scale. Consistent with this ideal, the benchmark t_b is here used as the value t would have if a collection of variables with the same frequencies as possessed by the variables in the cluster were a perfect Guttman scale. Correspondingly, we define t_c, the other benchmark, as the value t would have if a collection of variables with the same frequencies were completely independent of each other.

Design

Two different levels of within-cluster consistency, a_j, were chosen, corresponding to high and low consistency. These values, 3.0 and .5, are perhaps the extremes of range of values which might be expected in real data. A set of variables with $a_j = 3$ actually corresponds to a nearly perfect Guttman scale, while $a_j = .5$ is perhaps the lowest value that can occur for a set of variables to be recognizable as a cluster.

The parameters for the 24 variables, b_j, were selected so that the frequencies of cases scoring 1 approached 1/2. Data of this kind are good approximations of questionnaire or rating scale data, and are so common in the social sciences that we deemed it essential for BINCLUS to accurately recover these data sets before more stringent tests were undertaken.

Results and Discussion

There were three independent replications at the low consistency and two at the high. In all five cases, the clusters were identified exactly in every case. That is, each starting element added all the members of its own cluster before adding any from other clusters or any of the singletons. Furthermore, in most cases there seemed to be a definite drop in the value of the index after the last true member of the cluster was added, so that a user who was ignorant of the true identity and size of the clusters would have made the correct decisions concerning the extent of each cluster.

These results were most encouraging, particularly since the low a_j condition was felt to simulate the kind of consistency that is often found in real data, and real data of a poorish sort at that. Thus, it appears that where there are definite clusters or scales, the program with the q index will find them, even when there is considerable error in the data.

These results raised two questions. One concerned how well the program would perform with low-frequency data, since it will be remembered that the first study used variables with a

mean frequency of .5, although there was some variability. The second question had to do with the association index. We had started with the type of consistency index that seemed most appropriate for these data, but it was conceivable that the others would have worked as well.

Consequently, it was decided that a second, more elaborate study would be appropriate. It would use all four indices as the basis for clustering and would generate low-frequency as well as middle-frequency data. These data would be a closer approximation to those typically found in criminological investigations.

Experiment 2

Purposes of the Study

The basic aim of the second study was to examine the performance of BINCLUS using variables with low frequency. This is a more difficult condition because there is simply less information present when only five or two percent of the sample have positive scores on a variable. All of the association indices are based on joint-occurrence tables like Table 1, and when two variables have low frequencies, most of the observations are concentrated in the no-no cell, with the remainder scattered among the other three. Consequently, moving a single observation either to or from one of these three cells can have considerable effect on the index. For items that are stochastically independent of each other, one with a frequency of 5 percent and the

other with 2 percent, the expected number of cases with a score of 1 on both is .1 percent, one in a thousand. Nevertheless, this is often the type of data that one has in the criminal justice field, and it was necessary to investigate it.

This research was undertaken with the idea that q would have none of the defects that are felt to characterize the more familiar indices in dealing with this type of data. Nonetheless, the ones discussed here are familiar, more or less, and have wide use. Consequently, it was decided to try BINCLUS based on Pearson r (the Phi coefficient), Kuder-Richardson 20, which is so widely used as a measure of the internal consistency of a mental test, and the Goodman-Kruskal gamma. It was felt that the latter might behave much like q, since at least it exacted no penalty for unique relations.

Procedure

The general procedure differed only slightly from the original experiment. Again the clusters created according to the Birnbaum model contained 4, 6 and 8 variables. Together with the 6 unrelated variables there were a total of 24 as before and also 500 persons, as before. In addition to the high and low consistency data a moderate consistency condition was added. In the Birnbaum model the a_j parameter for low, moderate and high were .5, 1 and 3. Five sets of data were analyzed at each level of consistency.

The actual frequency distributions for the g_i and b_j remained normally distributed, but the frequency of the variables

was lowered from 50 percent to 10 percent. In the Birnbaum model the average parameter b_j was raised to 2.0 in the low and moderate consistency cases and to 1.5 in the high consistency cases.

Each of the 15 artificial data sets was submitted to the BINCLUS program using each of the four indices. Whereas in Experiment 1 the dependent variable had been the count of variables within a cluster recovered without error, in this study both the optimal percent correctly recovered and the percent recovered by using the largest gap rule were used.

Results

The mean proportions of correctly recovered clusters and a repeated measures analysis of variance for each of the dependent variables are shown in Tables 3 and 4.

Overall recovery rates as measured by the proportion of clusterns correctly placed was encouragingly high. Recovery for q, gamma, r and KR_{20}, respectively, are .91, .91, .92 and .92. Differences between the four indices, which can be seen to be rather minute, are in fact shown by analysis to be statistically nonsignificant.

Recovery as measured by the second criterion was lower, being .51, .48, .51 and .39 for the four indices in the same order as before. Here again the differences are not significant. The particularly low value for KR_{20} is due to its tendency to increase with cluster size, and therefore to interact badly with the largest gap rule. A rescaling of KR_{20} might improve its performance, but seemed unnecessary for the present investigation.

In other respects KR_{20} followed closely the behavior of Pearson's r and might be assumed to do so with regard to the stopping rule had such a rescaling taken place.

Discussion and Conclusions

The clustering method was again very successful according to the less stringent of the two criteria, even with low-frequency data which simulates actual criminal records. Over 90 percent of the variables were correctly included among the first k elements of a k-element cluster. The more stringent criterion that requires the identification of the boundary of a cluster by the largest gap was also supportive. About half the time, an objective rule that identifies the cluster boundary would identify as clusters exactly the k items that belong to them.

However, as a method for demonstrating the superiority of one index over another, Experiment 2 was certainly not successful. Performance of BINCLUS in recovering each of the clusters seemed to be quite robust to changes in the association index. The single departure from this uniformity was the failure of KR_{20} to function with the largest gap stopping rule as well as the other three indices. Even this departure can probably be minimized if there were sufficient incentive to rescale KR_{20} to fit the stopping rule.

The high level of recovery shown by the proportion of variables correctly placed is encouraging. This is especially true because only 500 cases were used in Experiment 2. At first

glance the roughly 50 percent recovery rates with the stopping rule are much less reassuring. However, it may be noted that if half the variables in an eight-member cluster, for example, lead to clusters which not only contain all the correct variables but only the correct variables, and 90 percent of the variables in every cluster are correctly located, the final decision regarding membership in the cluster, made after examination of all eight cluster histories is probably going to be correct or very nearly so.

In total then, the results of the first two experiments provide a solid basis for optimism regarding the performance of BINCLUS. It appears to be successful in recovering data which are frequently encountered in many kinds of social science research settings. Moreover, it does not require variables to be highly frequent, which is an asset for a potential tool in criminological studies.

Experiment 3

Purpose of the Study

In Experiments 1 and 2, the feasibility of analyzing prototypical criminological data with BINCLUS was demonstrated using artificial data. The similarity of performance using four different indices, however, did not answer the question of whether the indices would remain more or less interchangeable if analyses were undertaken with data having characteristics different from the ones used to generate the data of Experiment 2. The purpose of Experiment 3 was to provide some information

about the performance of BINCLUS with non-normal distributions.

Procedure

In many respects the details of Experiment 3 were similar to those of Experiment 2. The same dependent measures were used; the proportion of variables correctly placed and the number of perfectly recovered clusters using the stopping rule. The cluster structure was not changed. There was a cluster with four members, one with six and one with eight, as well as six outliers as in the previous experiments. The Birnbaum consistency parameter a_j was again set to .5, 1 and 3 to create low, moderate and highly consistent data. The overall frequencies of the variables were raised again to 50 percent as it had been in Experiment 1. However, the number of cases in the low, moderate and high consistency conditions was reduced to 100, 75 and 50, respectively, to study the program's performance with smaller samples and to push apart the four indices.

Unlike Experiments 1 and 2, here the shape of the distributions for the person characteristic q_i and the variables' characteristic b_j were altered so that in addition to normal distributions having means at zero and standard deviations of 1, there was also a condition which used rectangular distributions for both q_i and b_j; a condition which mixed a normal distribution of q_i with a rectangular distribution for b_j; and finally a condition which paired a normal q_i distribution with a bimodal b_j distribution. All the distributions had means of zero, but the bimodal distribution, instead of having a

standard deviation of 1 like all the others, consisted of two normal distributions whose means were -1 and +1 and whose standard deviations were .25.

As in Experiment 2, 5 data sets were created for each level of consistency under each type of distribution, so a total of 5 x 3 x 4 = 60 data sets were analyzed.

Results

Tables 5 and 6 display the repeated measures analysis of variance with the two dependent measures. Figure 1 displays graphically the means of the optimal dependent variable collapsed across cluster size. Cluster size was entered in the analysis as an independent variable in case the indices performed differentially depending on the cluster size. This did not happen. Cluster size did not have a significant interaction with index, but it did have a substantial main effect. Larger clusters were recovered better. The mean recovery rates for 4, 6 and 8 variable clusters are .71, .79 and .89. Here there is also a significant main effect for index used. The means for q, gamma, r and KR_{20} are .77, .78, .81 and .82, showing that the overall differences are small and in the direction opposite from what was expected. The interaction of the effect of index is significant with the level of cluster consistency (here confounded with the number of cases) and the type of distribution.

Figure 2 shows the corresponding means for the stopping rule recovery rate. Here there is also a significant, and larger, effect for the index used. The means for q, gamma, r and

KR_{20} are .24, .28, .21 and .07. Here they are in the expected direction, though with the exception of KR_{20} the differences are still small. The effect of index again has a significant interaction with level of cluster consistency/number of cases, but no significant interaction with type of distribution.

Overall Discussion and Conclusions

Overall Success

The overall outcome of this tryout with artificial data is that the clustering method works very well, but that it makes surprizingly little difference what index it is based on. With realistic data, success rates of correctly identifying clusters were as high as 100 percent. This rate of success can be degraded by making the data less reliable and by shifting the criterion of success to one that requires that it include correct identification of the number of elements in the cluster. If the frequencies of the simulated variables are reduced to near 10 percent, this has an effect. Also, if the sample sizes assumed are made smaller, again performance is affected. The internal consistency of the data also has an effect, but only when these other factors have had an opportunity to have an influence.

One of the key problems in the application of any clustering procedure is the correct identification of the number of elements in the cluster. Usually, the investigator is left with a subjective decision as to where to make the cut-off for a cluster or in how many clusters to accept. Here, we investigated

the effect of incorporating an objective rule: the end of a cluster is defined as the point at which the cluster-consistency index takes the largest drop. Thus, one makes a very stringent requirement of the procedure. Naturally, this reduces the success rate, but even so it is highly encouraging except under the most degraded conditions.

Comparison of Indices

The order of performance of the four indices was not consistent between the various designs and dependent measures. Because of the theoretical susceptibility of r and KR_{20} to disturbances caused by frequency differencies in the variables, it had been expected that q and perhaps gamma would consistently outperform them. This did not happen. In most instances the differences between the indices were small, too small to be important in empirical investigations. In case of a draw between indices when their performance is measured strictly by percent of item recovery, it might be argued that gamma is preferable because of its neutral attitude toward differences in frequency. An argument could be made for q, that by selecting variables of different frequencies, which gamma does not do, we may increase validity with some hypothetical external criteria. Aside from continuing a tradition of long misuse, no reason suggests itself for preferring r or KR_{20}.

The question of whether the indices actually <u>do</u> perform the same is open to argument. When considering the results of Experiment 3, the outcome is confused by multiple interactions

of the distributions of g_i and b_j, as well as the level of cluster consistency/number of cases factor. Most of the differences within cells of the first dependent variable analysis are on the order of .05 and the overall means have a range also equal to only .05. The results of the second dependent variable show much larger differences, and there is a good reason for weighting these results more heavily. The first dependent measure depends on the cluster histories being cut at an optimum level. This can be done because the cluster structure is known before hand. The second dependent variable is based on the cut point chosen automatically by the program, and therefore is more like the process which would occur in real data analysis. The largest consistent differences in the second dependent variable occur in the high cluster consistency cell of Experiment 3. Here the differences are as high as .59 in favor of q and gamma against r and KR_{20}. The overall means for this analysis differ only by .07, if KR_{20} is removed. KR_{20} worked particularly badly with the gap rule since it is an overall index rather than an average for only the candidate item. Nevertheless, if one obtained data with high consistency within clusters it would be an extremely bad idea to use r instead of q or gamma to find cluster boundaries. In all other cases the indices are more closely matched with the above mentioned exception of KR_{20}.

Comparison among the indices in the low frequency conditions, which are most relevant for analysis of criminological data, shows almost no differences at all. The range of the

means for the first dependent variable is .01 (.91 to .92). The range for dependent variable two is only .03 if we eliminate KR_{20}, with the means being .51, .48, .51 and .39 for q, gamma, r and KR_{20}, respectively.

The results of a clustering of typical data from a criminological study using police reports could probably be carried out equally well by q, gamma or r.

The overall performance of the method is encouraging in an absolute sense. Number of cases in the moderate frequency conditions had to be reduced to 100, 75 and 50 to prevent unvarying perfect recovery (with the first dependent variable at least). Performance of the low frequency conditions using 500 hypothetical cases may be revised upwards if one considers that the actual low frequency criminological data the program is to anlayze, contains 28,000 cases.

References

Bock, R. D. and Lieberman, M. Fitting a response model for n dichotomously scored items. Psychometrika, 1970, 35, 179-197.

Carroll, J. The nature of data or how to choose a correlation coefficient. Psychometrika, 1961, 26, 347-372.

Christoffersson, A. Factor analysis of dichotomized variables. Psychometrika, 1975, 40, 5-32.

Cliff, N. Test theory without true scores? Psychometrika, 1979, 44, 373-393.

Everitt, B. Cluster Analysis. London, Heienmann Educational Books Ltd., 1977.

Goodman, L. and Kruskal, W. Measures of Association for cross classifications. Journal of the American Statistical Association, 1954, 49, 732-764.

Green, B. Attitude measurement. In G. Linzey (Ed.) Handbook of Social Psychology. Cambridge: Addison-Welsey, 1954.

Guilford, J. and Fruchter, B. Fundamental Statistics in Psychology and Education. New York, McGraw-Hill, Inc., 1973.

Hardigan, J. Clustering Algorithms. New York, John Wiley and Sons, 1975.

Hofmann, R. Multiple hierarchical analysis. Applied Psychological Measurement, 1980, 4, 91-103.

Lord, F. and Novick, M. Statistical theories of mental test scores. Menlo Park, Addison-Wesley, 1968.

Nunnally, J. Psychometric Theory. New York, McGraw-Hill, Inc., 1967.

Sneath, P. and Sokal, R. Numerical Taxonomy, San Francisco, W. H. Freeman and Company, 1973.

Sokal, R. and Sneath, P. Principles of Numerical Taxonomy, London: Freeman, 1963.

Tryon, R. and Bailey, D. Cluster Analysis. New York: McGraw-Hill Book Co., 1970.

White, B. and Saltz, E. The measurement of reproducibility. Psychological Bulletin, 1957, 54, 81-99.

Table 3

Analysis of the Proportion of Variables Correctly Located within Clusters in Low Frequency Data

Source	SS	df	MS	F	α
Degree of Consistency (a_j)	3305.07778	2	1652.53889	2.52	0.0949
Cluster Size	3411.81111	2	1705.90556	2.60	0.0884
DC x CS	1500.98889	4	375.24722	0.57	0.6852
Error	23648.60000	36	656.90556		
Indices	66.44444	3	22.14815	0.72	0.5400
I x DC	163.85556	6	27.30926	0.89	0.5034
I x CS	79.12222	6	13.18704	0.43	0.8570
I x DM x CS	499.27778	12	41.60648	1.36	0.1968
Error	3305.80000	108	30.60926		

Means for the four indices:

q = .91
gamma = .91
r = .92
KR_{20} = .92

Table 4

Analysis of the Number of Clusters Exactly Recovered by the Automatic Stopping Rule in Low Frequency Data

Source	SS	df	MS	F	α
Degree of Consistency (a_j)	226.90000	2	113.45000	2.28	0.1450
Error	597,70000	12	49.80833		
Indices	48.85000	3	16.28333	1.79	0.1665
I x DC	255.90000	6	42.65000	4.69	0.0013
Error	327.50000	36	9.09722		

Mean proportions for the four indices:

q = .51
gamma = .48
r = .51
KR_{20} = .39

Table 5

Analysis of the Percent of Variables Correctly Located within Clusters in Moderate Frequency Data

Source	SS	df	MS	F	α
Distribution	5938.96111	3	1979.52037	2.11	0.1011
Degree of Consistency (a_j)	59913.90833	2	29956.95417	31.99	0.0000
Cluster Size	37446.17500	2	18723.08750	19.99	0.0000
D x DC	8304.88056	6	1384.14676	1.48	0.1898
D x CS	2537.68056	6	422.94676	0.45	0.8429
DC x CS	1677.71667	4	419.42917	0.45	0.7738
D x CD x CS	8973.02778	12	747.75231	0.80	0.6516
Error	134869.30000	144	936.59236		
Indices	2598.96111	3	866.32037	16.59	0.0000
I x D	1327.53889	9	147.50432	2.83	0.0031
I x DC	884.38056	6	147.39676	2.82	0.0105
I x CS	361.01389	6	60.16898	1.15	0.3310
I x D x DC	1428.91944	18	79.38441	1.52	0.0785
I x D x CS	857.48611	18	47.63812	0.91	0.5635
I x DC x CS	1062.79444	12	88.56620	1.70	0.0648
I x D x DC x CS	2214.10556	36	61.50293	1.18	0.2263
Error	22554.30000	432	52.20903		

Mean proportion for the four indices: q = .77; gamma = .78; r = .81; KR_{20} = .82

Mean proportion for three cluster sizes: 4 = .71; 6 = .79; 8 = .89

Table 6

Analysis of the Number of Clusters Exactly Recovered by the Automatic Stopping Rule in Moderate Frequency Data

Source	SS	df	MS	F	α
Distribution	48.98333	3	16.32778	0.63	0.6017
Degree of Consistency (a_j)	983.12500	2	491.56250	18.85	0.0000
D x DC	181.64167	6	30.27361	1.16	0.3429
Error	1251.90000	48	26.08125		
Indices	507.08333	3	169.02778	21.44	0.0000
I x D	82.85000	9	9.10556	1.17	0.3201
I x DC	469.84167	6	78.30694	9.93	0.0000
I x D x DC	187.92500	18	10.44028	1.32	0.1810
Error	1135.30000	144	7.88403		

Mean proportions for the four indices:

q = .24
gamma = .28
r = .21
KR_{20} = .07

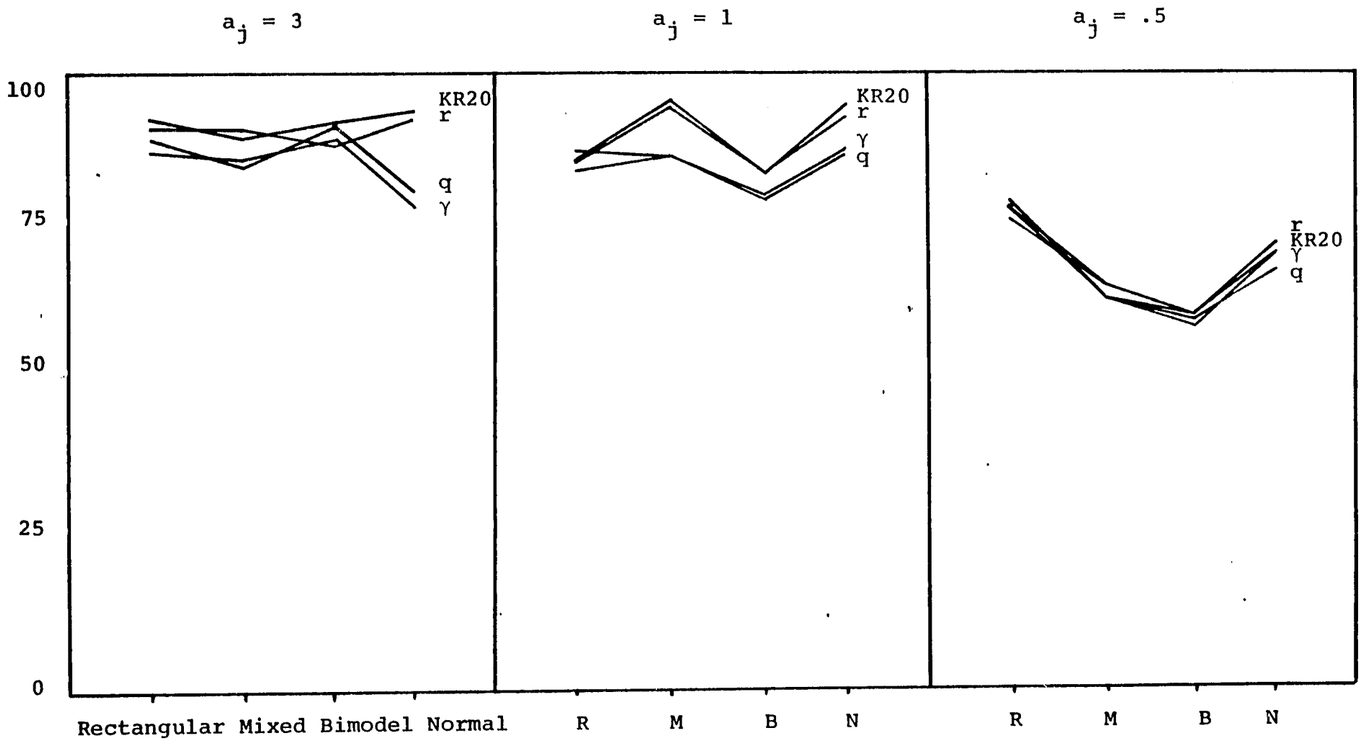

Figure 1

Proportion of Items Correctly Clustered in Moderate Frequency Data According to Distribution and Consistency (a_j)

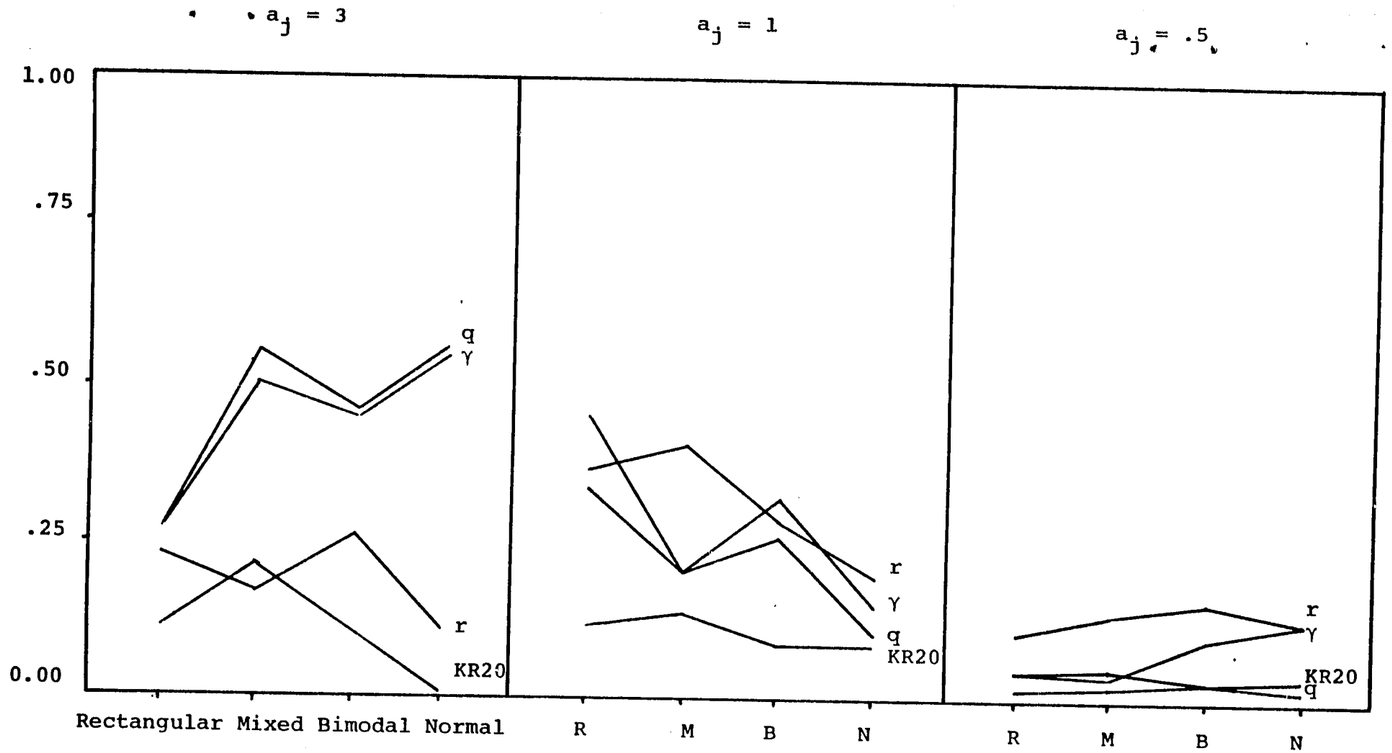

Figure 2

Proportion of Clisters Exactly Recovered from Moderate Difficulty Data Using the Automatic Stopping Rule According to the Degree of Cluster Consistancy (a_j) of the data and the Type of Distribution

END