REPRESENTATIVE SELECTION OF VARIABLES

by

Robert Arthur Bales

Thesis submitted to the Graduate Faculty of the

Virginia Polytechnic Institute

in candidacy for the degree of

MASTER OF SCIENCE

in

Statistics

October 1961

Blacksburg, Virginia

TABLE OF CONTENTS

INTRODUCTION

The classical and most frequent use of the discriminant function is for the purpose of classification, by obtaining a combined measurement such that the difference between two or more groups is as large as possible. The linear discriminant function for the special case of two groups was first discussed by R. A. Fisher (see [7]). A systematic generalization of the concept of discriminant functions was presented by S. N. Roy (see [11]), who utilized the discriminant function for the purpose of constructing tests of multivariate hypotheses. Discrimination between two groups has been widely utilized and is now discussed in many standard textbooks.

The discriminant function is a linear combination of the observable variables. We will speak of "observable" variables, if we can make direct measurements on them. All combinations into other functions, be they linear combinations or otherwise, will be referred to as "artificial" variables. In this sense, the linear discriminant function can be regarded as an artificial variable, for it represents a linear combination of the observable ones. Following S. N. Roy's approach, we may regard, as the discriminant function, that combination of the observable variables which produces the largest possible F-ratio between all groups under investigation. In this general notation, the discriminant function will be the eigenvector associated with the largest characteristic root of $E^{-1}H$, where E is a matrix of sums-of-squares and products due to error and H is a matrix of sums of squares and products due to a given hypothesis which specifies the groups. A more detailed exposition of this fact will be presented in Chapter I of this thesis.

The purpose of this thesis is to study the role of the discriminant function as a basis for selecting variables. Let us assume that we have k groups of experimental units, and on each experimental unit we make p observations. We would like to find that linear combination of the variables which produces a new variable such that the F-ratio between the k groups is a maximum. This linear combination of the variables is of course the discriminant function. We may regard this variable as an artificial variable constructed from the observed ones. We may then proceed to obtain the correlations of this artificial variable versus each of the observed variables. This vector of correlations will give us an idea of the proximity of the discriminant function to each of the observable variables. If, for example, we find a few correlations extremely high, say about .80, we can say that there is hardly any difference between the best discriminator between the k groups and those particular observed variables. We can then take the variables which have high correlations versus the discriminant function as almost equivalent in discriminatory power to that of the discriminant function itself. If, in further experimentation, we intend to reduce the number of variables on which we would like to perform measurements, we will of course select those variables which have the highest correlations (in absolute value) versus the best discriminator. This technique has been occasionally utilized for the selection of variables.

The best way to introduce the concept of <u>representative</u> selection is in the form of an example. Suppose we are interested in the difference between human beings of different races. We are not only interested in the physical characteristics but also, let us say, in their physiological differences, such as blood pressure, certain chemicals in the blood stream, respiration rate, etc. Let us assume, then, that we have three different racial groups and that we make 10 anthropological measurements, such as general height, weight, arm length, color of skin, etc., and that we also make 10 measurements on physiological characteristics such as blood pressure, content of calcium in the blood stream, etc. If we apply the previously mentioned discriminant function as a principle of selection, we will only include physical or anthropological variables in our selected set, because, naturally, the racial groups differ most strikingly in terms of physical characteristics. Thus, if we should perform a new set of experiments based only on the reduced set of variables, we would measure nothing but physical characteristics. This would completely obliterate any physiological measurements that we have previously conducted. It is in this context that the concept of <u>representative</u> selection enters. We may wish to select subsets out of the whole set of variables such that all the important common characteristics of the original set of variables are retained. It is thus necessary to classify the whole set of variables into certain subsets that belong together.

The classical method of finding subsets of variables such that they are classified under certain consistent principles of classification, is the so-called factor analysis (see [2] and [12]). In factor analysis, or, more generally, in Dependence Analysis, an attempt is made to explain the observed dependence among all the variables in terms of a few outside or artificial variables. These variates, incidentally, are not linear combinations of the original variables. This is a very striking difference between factor analysis and what is known as component analysis, which has no bearing on the present problem. The method of classification of subsets of variables is explained in detail in Chapter I, where the various preferred solutions of factor analysis have been explained and defined, and where the role of the "Simple Structure" as a classification principle is outlined. It is then shown that representative selection is performed in two stages. In the first stage, any group differences which may exist are eliminated in that we consider only the matrix of sums-of-squares and products due to error. From this matrix we construct a correlation matrix and on the basis of this correlation matrix we classify the variables into subsets.

In Chapter I the role of the discriminant function as a basis for the selection of variables, and the principle of representative selection by factor analysis is explained in detail. Chapter II presents a detailed demonstration study, where a factorial structure of the variables is known. It consists of a sampling experiment in which the variables have been arranged so as to form a structure in three artificial variables or factors. Chapter III presents a study of actual data which are an excerpt from a study on mentally retarded children, carried out in 1957-60 (see [13]). We selected three different sets of variables from these; intelligence measures, straight achievement measures, and measures of achievement gain in the course of one year. We were interested in discriminating between two groups; those children who stayed in the public school system, and those students who were transferred to special classes for mentally retarded children. Previous analysis of these daya showed that there was a rather striking interaction between the difference in school types and the age of the subjects, in that older children in the public education system performed considerably better than the older children that had been referred to special education classes. The difference was not nearly as great in the younger age groups. For that reason, we had to make three separate studies in investigating the discriminant functions and in finding the best representatives among the observed variables. A study was made for the children 11 years and younger, between 11 and 13 years of age, and for the group 13 years of age and older. The rather interesting and somewhat surprising findings of the studies are mentioned in Chapter III.

In the actual analysis of these data one must make use of algebraic advantages by representing matrices in the smallest possible form. The details of numerical evaluations in this case are presented in Chapter II.

For clarification it must be pointed out that the method of selecting variables on the basis of their correlations on the best discriminator does not insure that one will find subsets of 2, 3, or more variables which produce maximum F-ratios of all possible doublets, triplets, etc. No method for the selection of the best discriminating single variable can be used for finding, in general, the best two or three, for the "best" doublet may not even contain the best single variable. For illustration, let us consider the following example:

Let z_1 and z_2 be two uncorrelated variables, and let $z_3 = z_1 + z_2 + u$, where u is independent of z_1 and z_2 and has the same variance. Suppose we found, with z_1 and z_2 expressed in standard units, that $d = z_1 + z_2$ is the best discriminator between the groups under study in the sense that this combination of the three variables (z_1, z_2, z_3) would produce the highest F-ratio between groups. Thus, the correlations would be

	z_1	z_2	z_3	d
z_1	1.000	0.000	.577	.707
z_2	0.000	1.000	.577	.707
z_3	.577	.577	1.000	.816

Clearly, z_3 is the best single discriminating variable (it is, except for a relatively small error u, equal to the discriminant function itself, even though on the "partial" regression weights of d on z_3 given z_1 and z_2 it need not appear at all). However, the best <u>pair</u> of variables is z_1 and z_2 and does not even contain the best <u>single</u> variable z_3 . In fact, the multiple correlations of d on every possible pair are

$$R\ (d;\ z_1,\ z_2) = 1.000$$

$$R\ (d;\ z_1,\ z_3) = .866$$

$$R\ (d;\ z_2,\ z_3) = .866$$

so that z_3 is never a member of the best pair.

In the method proposed in this thesis the discriminant function is regarded as some unknown but real latent variable which produces maximum discrimination, and observable variables are selected in terms of their closeness to this ideal underlying variable. In this sense, it is natural to ask whether just one ideal variable would be sufficient to explain the discrimination, and it is for this reason that the representative selection has been proposed as an extension of this idea.

Chapter I: THEORY OF REPRESENTATIVE SELECTION

1.1 Notation

We shall present here some standard notation to facilitate reading:

(1.1.1) A' = transpose of a matrix A

(1.1.2) $\underline{x}$ = column vector; $\underline{x}'$ = row vector

(1.1.3) Σ = covariance matrix

(1.1.4) S = maximum-likelihood estimate of Σ

(1.1.5) $\tilde{R}$ = matrix of correlations in the population

(1.1.6) R = maximum-likelihood estimate of $\tilde{R}$ (sample correlations)

(1.1.7) F $(p \times k)$ = maximum-likelihood estimate of the matrix of correlations between observable and artificial variates, so-called "factor loadings"

(1.1.8) h_i^2 = sum-of-squares of the i'th row of F, called "communalities", which are squares of multiple correlations of each observed variable versus all artificial variables (factors)

(1.1.9) $R\,(\underline{z}|\underline{x})$ = matrix of all sample partial correlations in a set of variables, $\underline{z}$, given another set, $\underline{x}$

(1.1.10) D_a = diagonal matrix whose elements are $a_1, a_2, \ldots, a_p$

(1.1.11) E = matrix of sums-of-squares and products due to error

(1.1.12) H = matrix of sums-of-squares and products due to a hypothesis of no group differences.

1.2 Artificial Discriminator and Selection of Variables

In the sequel we will assume observation vectors taken from a multivariate normal population with common dispersion matrix Σ, and zero covariances between two observation vectors. Thus the observations for each experimental unit will constitute multivariate random variables from multivariate normal distributions, possibly with different mean vectors, but with equal variance-covariance matrices.

In univariate analysis, to test the difference between means of several groups we use the F-ratio MSH/MSE, where MSH is the mean-square between groups, and MSE is the mean-square error. In multivariate analysis the analogous test is that of equality of several mean vectors. Such a test is quite similar to the univariate case. One generalizes the sums-of-squares between groups, SSH, into a matrix H whose diagonal elements are the sums-of-squares between groups, for each of the p variables, and whose off-diagonal elements are corresponding sums-of-products between groups for pairs of variables. The sums-of-squares due to error, SSE, is generalized, by the same approach into a matrix E. Then, for the Union-Intersection test, discussed in [4] and [11], difference between mean vectors (or groups) is tested by finding the largest characteristic root of $E^{-1}H$, (see [8]). What is of interest here is the eignevector associated with this largest characteristic root of $E^{-1}H$, i.e., the vector $\underline{a}$ such that

(1.2.1) $\quad E^{-1}H\,\underline{a} = \lambda\,\underline{a}$.

This eigenvector $\underline{a}'$ is called the discriminant function between the groups; more specifically $\underline{a}'\underline{z}$, where $\underline{z}$ is the vector of observable variables, is that linear combination of $\underline{z}$ which produces maximum separation between mean effects, in that a univariate F test for the hypothesis of no group differences, performed on the artificial variable $u = \underline{a}'\underline{z}$, will produce the largest possible F-ratio.

As was stated, the discriminant function is an <u>artificial variable</u>, $u = \underline{a}'\underline{z}$. In other words, given a matrix of observations for n experimental units on p variables, i.e., a matrix Z (n x p), and a grouping principle which subdivides Z into groups Z_1 (n_1 x p), Z_2 (n_2 x p), ..., Z_k (n_k x p), we can obtain $E^{-1}H$ and hence a discriminant vector $\underline{a}'$. Then, if we form the vector $\underline{u} = Z\,\underline{a}$, i.e., a vector with n elements, the elements u_i may be regarded as the response on that artificial variable u which produces maximum separation between the k groups.

In this thesis we are interested in interpreting the artificial variable u . In particular, we would like to find those variables among the observable set $\underline{z}$ which are "closest" to the artificial variable u. If, therefore, one wants to select a subset of the $\underline{z}$ one would, for the purpose of optimum discrimination, choose the subset of those variables which are strongly correlated with u.

The rather general definition given here of a discriminant vector, $\underline{a}'$, in this case is clearly a sample quantity, since it is derived from sample quantities. A generalization of the univariate "Non-Centrality Parameter", in almost perfect analogy to the sample quantities given here (see [10]) leads to the definition of a corresponding vector $\underline{\alpha}'$. Suffice it to say, here, that $\underline{a}'$ is a maximum-likelihood estimate of $\underline{\alpha}'$ (for special cases see [1]).

To establish the association between observable variables $\underline{z}$, and the artificial variable u, we need the estimates of correlations between u and the original variable $\underline{z}$. Let, in the population,

(1.2.2) $\quad u = \underline{\alpha}'\underline{z}$

then

(1.2.3) $\quad \mathrm{var}\,(u) = \underline{\alpha}'\,\Sigma\,\underline{\alpha}$,

since $\mathrm{var}\,(\underline{z}) = \Sigma$. Also,

(1.2.4) $\quad \mathrm{cov}\,(u, \underline{z}') = \underline{\alpha}'\,\Sigma$,

(1.2.5) $\quad \mathrm{var}\,(z_i) = \sigma_{ii}$; hence,

(1.2.6) $\quad \mathrm{corr}\,(u, z_i) = \dfrac{1}{\sqrt{\underline{\alpha}'\,\Sigma\,\underline{\alpha}}}\,(\underline{\alpha}'\,\Sigma)_i \dfrac{1}{\sqrt{\sigma_{ii}}}$,

where $(\underline{\alpha}'\,\Sigma)_i$ is the i'th element of the row vector $\underline{\alpha}'\,\Sigma$.

Thus

$$(1.2.7) \qquad \text{corr}\,(u, \underline{z}') = \frac{1}{\sqrt{\underline{\alpha}' \, \Sigma \, \underline{\alpha}}} \, \underline{\alpha}' \, \Sigma \; D_{1/\sqrt{\sigma_{ii}}} \quad .$$

Now, the maximum-likelihood estimate of $\underline{\alpha}$ is $\underline{a}$, the maximum-likelihood estimate of Σ is $\frac{1}{n}$ E , and hence

$$(1.2.8) \qquad r\,(u, \underline{z}') = \frac{1}{\sqrt{\frac{1}{n}\,\underline{a}' \, E \, \underline{a}}} \, \frac{1}{n} \, \underline{a}' \, E \; D_{1/\sqrt{(e_{ii}/n)}}$$

$$= \frac{1}{\sqrt{\underline{a}' \, E \, \underline{a}}} \, \underline{a}' \, E \; D_{1/\sqrt{e_{ii}}} \quad .$$

The vector will be used for the purpose of selection of variables.

1.3 Representative Selection

Given a set of p observable variables, $z_1, z_2, \ldots, z_p$, we want to subdivide them into subsets. In other words, we want to find some principle of classification which enables us to put certain variables into a common class. This task is usually performed by Factor Analysis (see [2], [9], and [12]). The purpose of Factor Analysis is the study of dependence patterns in the variables. This is accomplished by seeking artificial variables which may explain the dependence among the observable variables. It is well known, in the theory of partial correlations, that dependence in a set of variables may be due to the presence of a common outside variable. The goal in Factor Analysis is to find one or more <u>artificial variables</u> which could explain the dependence of all the observable variates.* These artificial variables are described in terms of their correlations with the observable ones and these correlations have been called "factor loadings". The artificial variables, $\underline{x}$, must be determined in such a way that they "explain the dependence" in the set $\underline{z}$. Hence, if we can find $\underline{x}$ such that the hypothesis H_o: $\tilde{R}(\underline{z}|\underline{x}) = I$ is acceptable, the set $\underline{x}$ will serve our purpose. The solution of this problem for k factors ($k = 1, 2, \ldots$) is given by the relations (see [2])

$$(1.3.1) \qquad (R - F F') D_{1/1-h_i^2} F = F \; .$$

Any F satisfying this relation is a matrix of "factor loadings" which solves this problem, and h_i^2 is the sum-of-squares of the i'th row in F and also is the square of the multiple correlation (in the sample) of the artificial variables versus the i'th observable variable.

Suppose we choose $k = 1, 2, 3, \ldots$ and, for each choice of k, find an F satisfying <u>Equation (1.3.1)</u>. Given this F we can test the hypothesis $\tilde{R}(\underline{z}|\underline{x}) = I$ by a test of partial independence. It is, [1],

*There is the further attempt, analogous to the problem of fitting polynomials (see [2]), to keep the number of such artificial variables or factors as small as possible.

$$(1.3.2) \qquad -m \ln |R(\underline{z}|\underline{x})| \doteq \chi^2_f$$

where $m = n_e - \frac{2p+5}{6} - k$ and

$$f = p(p-1)/2 \quad ,$$

and n_e denotes the degrees of freedom due to error.

Ln $|R(\underline{z}|\underline{x})|$ can be reduced to the expression (see [2])

$$(1.3.3) \qquad \ln |R(\underline{z}|\underline{x})| = \ln |R| - \sum_{i=1}^{p} \ln (1 - h_i^2) + \ln |I - F' R^{-1} F|$$

and, actually, F as given in Equation (1.3.1) is chosen for each k in such a way that $|R(\underline{z}|\underline{x})|$ is a maximum (Maximum Determinant solution, identical to the Maximum-Likelihood solution of "Classical" Factor Analysis, see [9]). Using this χ^2-statistic for each k we can obtain a measure for the plausibiliy of the hypothesis that k factors are sufficient (see [2]). For each value of k, one evaluates such a probability α_k, and the sequence $\alpha_0, \alpha_1, \alpha_2, \ldots$, which is monotonically increasing, is called an indicator sequence (α_0 refers to the hypothesis of independence, i.e., zero factors). The decision on the proper number of factors is based on such a sequence. Fortunately, in all studies undertaken, (at Virginia Polytechnic Institute), the sequence has shown a sharp turning point with α_{k-1} very small and α_{k+1} very close to one, and α_k no smaller than 1/2 (see [5], and Chapter II and III of this Thesis), so that there can be no question as to the proper choice of the number of factors.

As implied before, the matrix F is not unique. For instance any orthogonal transformation on F will also satisfy Equation (1.3.1), i.e., let $F = F_o L$, where L is any orthogonal matrix, then

$F F' = F_o L L' F_o' = F_o F_o'$ and Equation (1.3.1) becomes

$$(1.3.4) \quad (R - F_o F_o') D_{1/1-h_i^2} F_o L = F_o L$$

$$= (R - F_o F_o') D_{1/1-h_i^2} F_o = F_o \quad .$$

Note that the h_i^2 are identical since they are the diagonal elements of $F F'$.

In the derivation of Equation (1.3.1) (see [2]), the artificial variables are assumed to be uncorrelated. An arbitrary orthogonal transformation of F will produce another F equally suitable for the problem at hand, as shown. But even an oblique transformation of F will produce a set of artificial variables (described in terms of their correlations with the observable ones) which still maximizes the determinant of $R(\underline{z}|\underline{x})$. In the attempt to find a representation of the artificial variables we are thus not limited to orthogonal transformations. Just as in the case of polynomial fitting the orthogonal polynomial representation is numerically the most elegant one, there is a corresponding solution ("Principal Axes") which is most desirable for a mathematical comparison of different studies.

This solution (see [11]) is an orthogonal transform $F\,L = P$ such that $P'P$ is diagonal. It is easy to see that L must be the matrix of eigenvectors of $F'F$.

The coefficients, γ_i, of the equation $y_i = \gamma_o + \gamma_1\,P_1\,(x_i) + \dots + \gamma_k\,P_k\,(x_i)$, where $P_j\,(x_i)$ denotes the j'th orthogonal (Tchebychev) polynomial of x, may be easy to obtain, mathematically, but they are certainly difficult to interpret, physically. For example, let y_i be the position of a mass falling in a vacumn, there the γ's in $y_i = \gamma_o + \gamma_1\,{}_1\,(t_i) + \gamma_2\,{}_2\,(t_i)$ are without meaning; but if we <u>rewrite</u> the right-hand side (without changing the function) in the representation $y_i = \alpha_o + \alpha_1\,t_i + \alpha_2\,t_i^2$, α_o can be interpreted as the position at time 0, α_1 can be interpreted as the initial speed, and α_2 as $\frac{1}{2}\,g$, where g is the acceleration due to gravity. This may serve as an illustration of the fact that different representations of the same function may serve different purposes.

The same is true of the representation of artificial variables in factor analysis. A transformation of the arbitrary matrix F will always result in a matrix whose elements are correlations of the artificial variables (combined and represented in some linear form) versus the observed ones.

For simplicity of interpretation, the Simple Structure concept has been proposed by Thurstone in [12]. The Simple Structure is obtained by finding individual transformation vectors such that they, when multiplied by F, yield other vectors with as many zero loadings as possible and just a few high loadings. If such a vector or vectors can be found, one can assign the variables to subsets. For we then have an artificial variable closely associated with a few of the observable ones and unrelated with the rest. Hence, the observable variables with the high loadings must contain some "factor" which does not influence the others. It is clear, then, that the variables with high loadings in such a vector should be assigned to a common subset or class. Ideally, when we have k factors, we would like to find just k such vectors of the simple structure. This would insure that all k artificial variables, which have been shown to be required to explain the dependence, are susceptible to interpretation. This ideal, however, is not always reached.

Once a classification of variables into subsets has been found, the problem of representative selection of variables for the purpose of discrimination can be solved by applying the formulae given in Section 1.2. Numerical short-cuts, especially in the evaluation of eigenvectors, will be presented in the next chapter.

Chapter II: DEMONSTRATION STUDY 1

2.1 Description of Data

This explicit demonstration has been constructed in order to investigate the possibilities of detecting underlying order in a sample. Fifteen sets of random normal numbers, 90 each, were obtained from Table A-23 of [6]. Table A-23 is a table of random normal numbers with mean 2 and variance 1, but each number taken from the table was augmented by 3 to produce random normal numbers with mean 5 and variance 1. From one given line in the table the first fifteen numbers were extracted to make one observation vector, and the first ninety lines were chosen to constitute ninety observation vectors.

These fifteen variables were denoted as:

Set 1: x_1

Set 2: x_2

Set 3: x_3

Sets 4-15: $u_1 - u_{12}$.

These sets were further subdivided into 3 groups of 30 "experimental units" each, i.e., each group consisted of 30 observation vectors. Group effects were introduced into the variable x_1, x_2, x_3, u_3, u_5, and u_8, in the following manner:

In group 1, each value of the set x_1 was augmented by 6, x_2 by 0, x_3 by 2, u_3 by 4, u_5 by 2, and u_8 by 6; the remaining sets were unchanged.

In group 2, each value of the set x_1 was augmented by 3, x_2 by 2, x_3 by 0, u_3 by 2, u_5 by 1, and u_8 by 6; the remaining sets were unchanged.

In group 3, each value of the set x_1 was augmented by 0, x_2 by 4, x_3 by 1, u_3 by 0, u_5 by 0, and u_8 by 0; the remaining sets were unchanged.

Now using these augmented variables, a new set of variables was constructed as follows:

$$
\begin{aligned}
z_1 &= 3\,x_1 + u_1 \\
z_2 &= 5\,x_1 + u_2 \\
z_3 &= x_1 + u_3 \\
z_4 &= 2\,x_1 + u_4 \\
z_5 &= 6\,x_2 + u_5 \\
z_6 &= 3\,x_2 + u_6 \\
z_7 &= 2\,x_2 + u_7 \\
z_8 &= x_2 + u_8 \\
z_9 &= x_3 + u_9 \\
z_{10} &= 4\,x_3 + u_{10} \\
z_{11} &= x_1 + x_2 + u_{11} \\
z_{12} &= x_2 + 6\,x_3 + u_{12}
\end{aligned}
$$

$z_1, z_2, \ldots, z_{12}$ will be used as the "observable" variables for the study. It will be noted that this leads to a demonstration of a three-factor study, with x_1, x_2, and x_3 as common factors and the uncorrelated u's as specifics. (The population parameters of the z variables are presented in Table 2.1.1.) Thus, the analysis will be made on the z variables only, and the structure of these variables will be assumed unknown. This will enable us to compare the results of the analysis with the actual underlying structure later on. The technique of finding the unknown structure will be the usual Factor Analysis method.

2.2 Preparation of Data

The matrix of sums-of-squares and sums-of-products within groups E is given in Table 2.2.1 and E^{-1} is given in Table 2.2.2. The construction of the z variables, the matrix E, and its inverse E^{-1} were obtained on the I. B. M. 650 electronic computer by use of the Revised General Multiple Regression System (06.2.008). The matrix of estimated correlations R was obtained in the following manner: Let the (i, j)th element of E be e_{ij}, then the estimated correlation of the ith and jth variable would be

$$r_{ij} = \frac{e_{ij}}{\sqrt{e_{ii}}\,\sqrt{e_{jj}}} .$$

Thus,

$$(2.2.1) \qquad R = D_{1/\sqrt{e_{ii}}}\, E\, D_{1/\sqrt{e_{ii}}}$$

where $D_{1/\sqrt{e_{ii}}}$ is a diagonal matrix with diagonal elements

Table 2.1.1

Population Means and Variances*

	Group I	Group II	Group III
z_1	38, 10	29, 10	20, 10
z_2	60, 26	45, 26	30, 26
z_3	20, 2	15, 2	10, 2
z_4	27, 5	21, 5	15, 5
z_5	37, 37	48, 37	59, 37
z_6	20, 10	26, 10	32, 10
z_7	15, 5	19, 5	23, 5
z_8	16, 2	15, 2	14, 2
z_9	12, 2	10, 2	11, 2
z_{10}	33, 17	25, 17	29, 17
z_{11}	21, 3	20, 3	19, 3
z_{12}	52, 38	42, 38	50, 38

*An entry in the table of μ, σ^2 means that the variable is distributed normal with mean μ and variance σ^2 .

Table 2.2.1

Matrix E

	1	2	3	4
1	783.2361	1153.8684	244.1520	470.9815
2	1153.8684	1987.4942	407.7200	771.9945
3	244.1520	407.7200	161.2637	149.1448
4	470.9815	771.9945	149.1448	389.9145
5	216.7007	227.3985	6.0731	87.0861
6	106.3894	117.4361	2.6563	43.3284
7	19.2648	23.9384	- 21.3434	23.2976
8	19.8597	20.7101	- 18.6569	13.7184
9	- 31.2640	- 38.0216	- 10.3488	5.7388
10	- 97.9325	- 152.7416	- 16.5333	- 20.2888
11	243.8801	416.7521	71.2785	169.6193
12	- 81.7115	- 113.3096	7.3703	12.5446

	5	6	7	8
1	216.7007	106.3894	19.2648	19.8597
2	227.3985	117.4361	23.9384	20.7101
3	6.0731	2.6563	- 21.3434	- 18.6569
4	87.0861	43.3284	23.2976	13.7184
5	2924.1632	1465.9236	881.6001	406.0168
6	1465.9236	802.0511	440.4736	206.1355
7	881.6001	440.4736	372.4585	136.2462
8	406.0168	206.1355	136.2462	148.1918
9	7.2575	- 4.1125	10.4524	7.6397
10	- 253.9752	- 142.1466	- 62.2690	- 39.6836
11	494.6827	251.1353	153.8364	83.1061
12	173.6893	68.0215	93.9487	17.6623

Table 2.2.1 (Cont.)

	9	10	11	12
1	- 31.2640	- 97.9325	243.8801	- 81.7115
2	- 38.0216	- 152.7416	416.7521	- 113.3096
3	10.3488	- 16.5333	71.2785	7.3703
4	5.7388	- 20.2888	169.6193	12.5446
5	7.2575	- 253.9752	494.6827	173.6893
6	- 4.1125	- 142.1466	251.1353	68.0215
7	10.4524	- 62.2690	153.8364	93.9487
8	7.6397	- 39.6836	83.1061	17.6623
9	155.3080	250.3869	- 2.8052	386.2087
10	250.3869	1078.1841	- 50.6908	1484.6407
11	- 2.8052	- 50.6908	240.5619	50.9893
12	386.2087	1484.6407	50.9893	2415.9008

Table 2.2.2

Inverse of E (x 100) 100 E^{-1}

	1	2	3	4
1	1.048036	- .467815	- .114314	- .338727
2	- .467815	.563363	- .294057	- .322011
3	- .114314	- .294057	1.442965	.137537
4	- .338727	- .322011	.137537	1.293669
5	- .151980	.075709	- .048972	.093598
6	.065273	- .031236	.014043	- .007524
7	.205905	- .031564	.137828	- .196459
8	- .017160	.012133	.216205	- .009795
9	.115241	- .022437	- .064943	- .170075
10	- .070300	.025528	.134557	- .016213
11	.117531	- .290285	.078825	- .114159
12	.039160	.003720	- .121298	.007088

	5	6	7	8
1	- .151980	.065273	.205905	- .017160
2	.075709	- .031236	- .031564	.012133
3	- .048972	.014043	.137828	.216205
4	.093598	- .007524	- .196459	- .009795
5	.547898	- .756221	- .311455	- .078865
6	- .756221	1.507099	.042809	- .043135
7	- .311455	.042809	1.069861	- .102457
8	- .078865	- .043135	- .102457	1.191245
9	- .056918	.080648	.031569	- .088476
10	.050679	.026324	.103608	.002519
11	- .123679	- .031557	- .070042	- .203996
12	- .026777	- .017077	- .080821	.018440

Table 2.2.2 (Cont.)

	9	10	11	12
1	.115241	- .070300	.117531	.039160
2	- .022437	.025528	- .290285	.003720
3	- .064943	.134557	.078825	- .121298
4	- .170075	- .016213	- .114159	.007088
5	- .056918	.050679	- .123679	- .026777
6	.080648	.026324	- .031557	- .017077
7	.031569	.103608	- .070042	- .080821
8	- .088476	.002519	- .203996	.018440
9	1.129534	- .081994	.088817	- .127282
10	- .081994	.880481	.098820	- .540000
11	.088817	.098820	1.301669	- .097665
12	- .127282	- .540000	- .097665	.402893

$1/\sqrt{e_{ii}}$. The calculation of R was accomplished by the Interpretive Matrix Operations (05.2.002) on the I. B. M. 650. The matrix R is given in Table 2.2.3, and R^{-1} in Table 2.2.4.

2.3 Factor Analysis

The Factor Analysis performed in this study was done in order to determine the structure of this sampling problem. The rotation to the Simple Structure will enable us to divide the variables into representative sets. Once this structure has been determined, the discriminant function for each set may be calculated and in turn one may find the correlations of the observable variables versus the discriminant function (artificial variable). These correlations will then be the basis for ordering the observable variables.

The complete Factor Analysis was computed on the I. B. M. 650 by Factor Analysis programs (6.6.021) of [3].

Table 2.3.1 represents an improved centroid solution which was used as input for the maximum-likelihood iterations and for the purpose of making a preliminary decision regarding the number of factors. This preliminary test is given in Table 2.3.2.

Table 2.2.3

Matrix of Correlations R

	1	2	3	4
1	1.000000	.924819	.686982	.852261
2	.924819	1.000000	.720179	.876954
3	.686982	.720179	1.000000	.594779
4	.852261	.876954	.594779	1.000000
5	.143190	.094327	.008844	.081557
6	.134230	.093014	.007386	.077479
7	.035668	.027823	- .087087	.061135
8	.058293	.038161	- .120687	.057070
9	- .089640	- .068435	- .065392	.023321
10	- .106570	- .104342	- .039650	- .031291
11	.561845	.602714	.361890	.553831
12	- .059401	- .051612	.011808	.012925

	5	6	7	8
1	.143190	.134230	.035668	.058293
2	.094327	.093014	.027823	.038161
3	.008844	.007386	- .087087	- .120687
4	.081557	.077479	.061135	.057070
5	1.000000	.957215	.844757	.616781
6	.957215	1.000000	.805898	.597992
7	.844757	.805898	1.000000	.579927
8	.616781	.597992	.579927	1.000000
9	.010769	- .011652	.043459	.050358
10	- .143035	- .152858	- .098262	- .099278
11	.589811	.571733	.513933	.440157
12	.065348	.048866	.099040	.029519

Table 2.2.3 (Cont.)

	9	10	11	12
1	- .089640	- .106570	.561845	- .059401
2	- .068435	- .104342	.602714	- .051612
3	- .065392	- .039650	.361890	.011808
4	.023321	- .031291	.553831	.012925
5	.010769	- .143035	.589811	.065348
6	- .011652	- .152858	.571733	.048866
7	.043459	- .098262	.513933	.099040
8	.050358	- .099278	.440157	.029519
9	1.000000	.611883	- .014513	.630501
10	.611883	1.000000	- .099533	.919889
11	- .014513	- .099533	1.000000	.066885
12	.630501	.919889	.066885	1.000000

Determinant of R = .0000033889622

Table 2.2.4

Matrix R^{-1}

	1	2	3	4
1	8.208512	- 5.836711	- .406284	- 1.871866
2	- 5.836711	11.196745	- 1.664739	- 2.834741
3	- .406284	- 1.664739	2.326971	.344873
4	- 1.871866	- 2.834741	.344873	5.044218
5	- 2.299995	1.825134	- .335983	.999417
6	.517384	- .394414	.050189	- .042063
7	1.112115	- .271547	.337777	- .748696
8	- .058538	.065904	.334230	- .023539
9	.401938	- .124661	.102772	- .418528
10	- .645977	.373719	.561058	- .105179
11	.510165	- 2.007187	.155259	- .349638
12	.538625	.081499	- .757098	.068846

	5	6	7	8
1	- 2.299995	.517384	1.112115	- .058538
2	1.825134	- .394414	- .271547	.065904
3	- .335983	.050189	.337777	.334230
4	.999417	- .042063	- .748696	- .023539
5	16.020433	-11.580732	- 3.250564	- .517588
6	-11.580732	12.087967	.234184	- .150351
7	- 3.250564	.234184	3.984793	- .240745
8	- .517588	- .150351	- .240745	1.765369
9	- .383667	.284755	.075933	- .134266
10	.899803	.244841	.656583	.010051
11	- 1.037676	- .138248	- .209648	- .385151
12	- .711569	- .237859	- .766676	.110353

Table 2.2.4 (Cont.)

	9	10	11	12
1	.401938	- .645977	.510165	.538625
2	- .124661	.373719	- 2.007187	.081499
3	.102772	.561058	.155259	- .757098
4	- .418528	- .105179	- .349638	.068846
5	- .383667	.899803	- 1.037676	- .711569
6	.284755	.244841	- .138248	- .237859
7	.075933	.656583	- .209648	- .766676
8	- .134266	.010051	- .385151	.110353
9	1.754263	- .335519	.171683	- .779670
10	- .335519	9.493215	.503281	- 8.715259
11	.171683	.503281	3.131309	- .744553
12	- .779670	- 8.715259	- .744553	9.733502

Table 2.3.1

Improved Centroid Solution for F

	1	2	3	4	5
1	.639272	.353586	.617691	-.031142	-.019289
2	.648319	.344169	.651531	.062182	.065902
3	.417864	.200558	.565298	-.177402	.026378
4	.630414	.242303	.581188	.203596	-.109026
5	.668936	.286984	-.644712	-.178239	-.088879
6	.638772	.295459	-.624540	-.192338	-.055062
7	.564108	.200344	-.626736	.016578	-.024671
8	.434837	.176577	-.493017	.199836	.023676
9	.258480	-.628913	-.044823	.151889	-.136842
10	.245689	-.911991	.077838	-.057242	.036091
11	.765070	.316069	-.026707	.096924	.180986
12	.432299	-.875119	-.033021	-.094667	.100788

Table 2.3.2

Tests of Partial Independence (see (1.3.2))

(Preliminary)

	χ^2_{66}	Probability
Independence	1034.89	$< 10^{-6}$
1 Factor	977.05	$< 10^{-6}$
2 Factors	698.43	$< 10^{-6}$
3 Factors	114.34	.0001
4 Factors	62.56	.60
5 Factors	49.80	.93

Clearly, two factors are insufficient to explain the observed dependence. The decrease in the χ^2-value is rather marked but, according to the preliminary test, three factors are not sufficient. Four factors are clearly sufficient. It must be noted, however, that the χ^2-value will be smaller, and hence the associated probability larger, once the maximum-likelihood solution is available. Following the preliminary decision, 4 factors were used for the iteration into a maximum-likelihood solution. The procedure was stopped when all communalities agreed to four places. This required 35 iterations. The test for partial independence was repeated yielding, for 4 factors, a χ^2 of 13.105 (66 d.f.) corresponding to a normal curve equivalent of 6.33, which denotes an extremely high probability ($> 1-10^{-9}$). There may thus be a possibility that three factors would be quite sufficient. The maximum-likelihood solution for three factors was also established (after 45 iterations starting from centroid) and, finally, also that for 2 factors (after 10 iterations starting from the principal-axes form of the three-factor solution). The χ^2-values for 3 and 2 factors, respectively, were $\chi^2 = 28.702$ (66 d.f.), normal curve equivalent 3.87, corresponding to a probability of .99995 ; $\chi^2 = 247.58$ (66 d.f.), normal curve equivalent -10.81, corresponding to a probability less than 10^{-20}. Hence the final indicating sequence (see [2]) is, for the probability of k factors being sufficient,

$$\alpha_2 < 10^{-20}$$

$$\alpha_3 = .99995$$

$$\alpha_4 > 1 - 10^{-9} \quad ,$$

and there can be no doubt about the fact that three factors is the correct number. Thus, the test correctly identifies the known underlying degree of dependence. The (arbitrary) maximum-likelihood solution F for three factors is shown in Table 2.3.3. This was rotated into the Principal-Axes form which is shown in Table 2.3.4. For interpretation of results, a Simple Structure (see [12]) solution was sought which is shown in Table 2.3.5. If there were no sampling errors the Simple Structure would look as shown in Table 2.3.6. The values of this theoretical structure can be obtained by the following argument:

Consider the variable, $z_i = a\, x_1 + b\, x_2 + c\, x_3 + u_i$, then

(2.3.1) $$h_i^2 \text{ (communality)} = \frac{\text{var } (z_i - u_i)}{\text{var } (z_i)} \quad .$$

Now, the x_1 loading is $a\, k$, x_2 loading $b\, k$, and x_3 loading $c\, k$, where

(2.3.2) $$(a\, k)^2 + (b\, k)^2 + (c\, k)^2 = h_i^2 \quad ;$$

e.g., consider $z_{12} = x_2 + 6\, x_3 + u_{12}$:

Table 2.3.3

Maximum-Likelihood Solution

F Matrix after 45 Iterations (3 Factor)

	1	2	3
1	.635837	.353673	.594324
2	.640844	.347614	.665759
3	.450229	.195895	.538746
4	.603770	.255958	.605164
5	.681370	.287370	-.663668
6	.653979	.292243	-.643477
7	.565200	.194770	-.607847
8	.410116	.183341	-.433747
9	.240975	-.596202	-.024891
10	.254358	-.919182	.079711
11	.754218	.301190	.012109
12	.434745	-.883480	-.035051

Table 2.3.4

Principal Axes (3 Factor)

	I	II	III	h_i^2
1	.762427	.545839	-.057958	.882594
2	.768341	.615496	-.074623	.974752
3	.521373	.498714	-.103841	.531328
4	.688834	.552083	-.130340	.796276
5	.694425	-.710350	.021893	.987302
6	.673851	-.686854	.036257	.927157
7	.551983	-.649625	-.012881	.726864
8	.420767	-.460643	.026622	.389946
9	-.070372	-.101467	-.631584	.414145
10	-.204377	-.026255	-.934606	.915947
11	.806310	-.048369	-.085040	.659707
12	-.035835	-.157335	-.971973	.970769

Table 2.3.5

Simple Structure 3 Factors

Var.	A	B	C	h_i^2	Belonging to Subset
1	.927	.062	-.041	.8826	A
2	.981	.011	-.030	.9748	A
3	.729	-.059	.025	.5313	A
4	.890	.005	.036	.7963	A
5	.019	.989	-.008	.9873	B
6	.017	.959	-.023	.9272	B
7	-.037	.849	.031	.7269	B
8	-.011	.624	-.016	.3899	B
9	-.017	-.010	.641	.4141	C
10	-.014	-.174	.945	.9159	C
11	.563	.544	.034	.6597	A and B
12	.026	.031	.982	.9708	C

Table 2.3.6

Expected Simple Structure (3 Factors)

	x_1	x_2	x_3	h_i^2
1	.949	.000	.000	.900
2	.981	.000	.000	.962
3	.707	.000	.000	.500
4	.894	.000	.000	.800
5	.000	.986	.000	.973
6	.000	.949	.000	.900
7	.000	.894	.000	.800
8	.000	.707	.000	.500
9	.000	.000	.707	.500
10	.000	.000	.970	.941
11	.577	.577	.000	.667
12	.000	.162	.973	.974

$$h_{12}^2 = 37/38 \quad ,$$

$$k^2 + 36\,k^2 = 37/38 \;, \quad \text{or} \quad k^2 = 1/38 \quad ,$$

$$k = 1/\sqrt{38} = .162 \quad .$$

Hence the x_1 loading is 0, x_2 loading .162 and x_3 loading .973 (6 x 1.62).

A comparison of Table 2.3.5 (observed Simple Structure) with Table 2.3.6 (theoretical Simple Structure based upon the population parameters) shows the good agreement of all values, and hence demonstrates the usefulness of the method of analysis, even for a comparatively small sample.

Incidentally, if we had taken the overfactored study (4 factors) as a starting point, the same result would have shown in the Simple Structure. Table 2.3.7 shows the Simple Structure formed from the four factor solution. A fourth overdetermined plane could not be found.

As expected the Simple Structure produced three representative sets and they are as follows:

Set I - Variables 1, 2, 3, 4, and 11

Set II - Variables 5, 6, 7, 8, and 11

Set III - Variables 9, 10, and 12 .

2.4 Discriminatory Analysis

The discriminant function $\underline{a}$ is given by the expression

$$E^{-1}\,H\,\underline{a} = \lambda\,\underline{a} \quad , \tag{2.4.1}$$

Table 2.3.7

Simple Structure   4 Factors

Var.	A	B	C	Belonging to Subset
1	.914	.066	-.037	A
2	.970	.007	-.032	A
3	.695	-.040	.036	A
4	.902	-.008	.029	A
5	.004	.992	.000	B
6	.005	.952	-.016	B
7	-.023	.828	.025	B
8	.017	.596	-.026	B
9	-.007	-.017	.637	C
10	-.017	-.172	.947	C
11	.577	.522	.026	A and B
12	.022	.029	.978	C

and the solution for $\underline{a}$ is the eigenvector associated with the largest characteristic root of $E^{-1}H$, where E^{-1} is the matrix given in Table 2.2.2 and H is the matrix for the hypothesis of no group differences. Hence the elements must be sums-of-squares and products "between groups", or

$$(2.4.2) \qquad h_{ij} = \sum_{r=1}^{3} n_r(\bar{x}_r^{(i)} - \bar{\bar{x}}^{(i)})(\bar{x}_r^{(j)} - \bar{\bar{x}}^{(j)}) \quad ,$$

where $\bar{x}_r^{(p)}$ is the mean of the p'th variable in the r'th group, and $\bar{\bar{x}}^{(p)}$ is the mean of the p'th variable over all 3 groups.

Now, for ease of calculation let $B\,B' = H$, where the (i, r)'th element of B is $(\bar{x}_r^{(i)} - \bar{\bar{x}}^{(i)})\sqrt{n_r}$, and n_r is the size of group r. Thus the matrix B in Set I mentioned above would be:

$$\sqrt{30}\begin{bmatrix} (\bar{x}_1^{(1)} - \bar{\bar{x}}^{(1)}) & (\bar{x}_2^{(1)} - \bar{\bar{x}}^{(1)}) & (\bar{x}_3^{(1)} - \bar{\bar{x}}^{(1)}) \\ (\bar{x}_1^{(2)} - \bar{\bar{x}}^{(2)}) & (\bar{x}_2^{(2)} - \bar{\bar{x}}^{(2)}) & (\bar{x}_3^{(2)} - \bar{\bar{x}}^{(2)}) \\ (\bar{x}_1^{(3)} - \bar{\bar{x}}^{(3)}) & (\bar{x}_2^{(3)} - \bar{\bar{x}}^{(3)}) & (\bar{x}_3^{(3)} - \bar{\bar{x}}^{(3)}) \\ (\bar{x}_1^{(4)} - \bar{\bar{x}}^{(4)}) & (\bar{x}_2^{(4)} - \bar{\bar{x}}^{(4)}) & (\bar{x}_3^{(4)} - \bar{\bar{x}}^{(4)}) \\ (\bar{x}_1^{(11)} - \bar{\bar{x}}^{(11)}) & (\bar{x}_2^{(11)} - \bar{\bar{x}}^{(11)}) & (\bar{x}_3^{(11)} - \bar{\bar{x}}^{(11)}) \end{bmatrix}$$

using variables 1, 2, 3, 4, 11. These calculations were further eased by utilizing the following relations:

$\bar{x}_r^{(i)} = \frac{T_r^{(i)}}{30}$ where $T_r^{(i)}$ is the total of the i'th variable in group r ; $\bar{\bar{x}}^{(i)} = \frac{G^{(i)}}{90}$ where $G^{(i)}$ is the total of the i'th variable over all three groups.

Now, by Equation (2.4.1),

(2.4.3) $E^{-1} B B' \underline{a} = \lambda \underline{a}$ and

(2.4.4) $B' E^{-1} B [B' \underline{a}] = \lambda[B' \underline{a}]$.

Now let $B' \underline{a} = \underline{u}$, hence

(2.4.5) $B' E^{-1} B \underline{u} = \lambda \underline{u}$.

We find the largest characteristic root of the small symmetric matrix $B' E^{-1} B$ and the eigenvector associated with the root is $\underline{u}$. (Note that $E^{-1} H$ would be non-symmetric, 12 x 12, whereas $B' E^{-1} B$ is symmetric, 3 x 3.) Then from Equation (2.4.5),

(2.4.6) $E^{-1} B B' E^{-1} B \underline{u} = \lambda E^{-1} B \underline{u}$ or

(2.4.7) $E^{-1} H [E^{-1} B \underline{u}] = \lambda [E^{-1} B \underline{u}]$, hence

(2.4.8) $\underline{a} = E^{-1} B \underline{u}$

and the desired discriminant function is solved.

This procedure was followed in obtaining the discriminant function for each of the three sets derived from the Simple Structure and also for the set of all twelve variables taken together. The task was performed as follows:

(a) The matrix B computed

(b) The matrix $E^{-1} B$ and $B' E^{-1} B$ computed using the Matrix Interpreter

(c) Characteristic roots and associated eigenvectors of $B' E^{-1} B$ were obtained

(d) The largest characteristic root and associated eigenvector $\underline{u}$ was chosen

(e) Using $\underline{u}$ and $E^{-1} B$ from above $E^{-1} B \underline{u}$ was computed.

The final step gives $\underline{a}$.

<u>Set I</u> (Variables 1, 2, 3, 4, 11)

The matrix B is given in <u>Table 2.4.1</u> and E in <u>Table 2.4.2</u>. The discriminant function is

$$\underline{a}' = [.0828, .2670, 1.6459, .3809, -.9859]$$

or in terms of the z_i's ,

$$\underline{a}'\underline{z} = [828\ z_1 + 2{,}670\ z_2 + 16{,}459\ z_3 + 3{,}809\ z_4 - 9{,}859\ z_{11}].$$

<u>Set II</u> (Variables 5, 6, 7, 8, 11)

The matrix B is given in <u>Table 2.4.3</u> and E in <u>Table 2.4.4</u>. The discriminant function is

$$a' = [.0946, .7880, .8568, -2.0283, -1.1757]$$

or in terms of the z_i's ,

$$\underline{a}'\underline{z} = [946\ z_5 + 7{,}880\ z_6 + 8{,}568\ z_7 - 20{,}283\ z_8 - 11{,}757\ z_{11}] .$$

<u>Set III</u> (Variables 9, 10, 12)

The matrix B is given in <u>Table 2.4.5</u> and E in <u>Table 2.4.6</u>. The discriminant function is

Table 2.4.1

Matrix B Set I (Variables 1, 2, 3, 4, 11)

	1	2	3
1	49.8840	1.2101	-51.0941
2	81.3360	4.1348	-85.4709
3	25.6413	2.4560	-28.0973
4	31.8943	2.7882	-34.6825
11	4.5844	1.4907	- 6.0750

Table 2.4.2

Matrix E Set I (Variables 1, 2, 3, 4, 11)

	1	2	3	4	11
1	783.2361	1153.8684	244.1520	470.9815	243.8801
2	1153.8684	1987.4942	407.7200	771.9945	416.7521
3	244.1520	407.7200	161.2637	149.1448	71.2785
4	470.9815	771.9945	149.1448	389.9145	169.6193
11	243.8801	416.7521	71.2785	169.6193	240.5619

Table 2.4.3

Matrix B Set II (Variables 5, 6, 7, 8, 11)

	1	2	3
5	-57.6623	4.8951	52.7672
6	-31.9302	2.4607	29.4696
7	-21.7481	2.2249	19.5232
8	5.5677	.2918	- 5.8595
11	4.5844	1.4907	- 6.0750

Table 2.4.4

Matrix E Set II (Variables 5, 6, 7, 8, 11)

	5	6	7	8	11
5	2924.1632	1465.9236	881.6001	406.0168	494.6827
6	1465.9236	802.0511	440.4736	206.1355	251.1353
7	881.6001	440.4736	372.4585	136.2462	153.8364
8	406.0168	206.1355	136.2462	148.1918	83.1061
11	494.6827	251.1353	153.8364	83.1061	240.5619

Table 2.4.5

Matrix B Set III (Variables 9, 10, 12)

	1	2	3
9	4.1024	- 6.1467	2.0443
10	15.4846	-18.4680	2.9834
12	13.0248	-29.1103	16.0855

Table 2.4.6

Matrix E Set III (Variables 9, 10, 12)

	9	10	12
9	155.3080	250.3869	386.2087
10	250.3869	1078.1841	1484.6407
12	386.2087	1484.6407	2415.9008

$$a' = [.1671,\ .5778,\ -.3398]$$

or in terms of the z_i's ,

$$\underline{a}'\underline{z} = [1{,}671\ z_9 + 5{,}778\ z_{10} - 3{,}398\ z_{12}]\ .$$

Set of all variables

The matrix B is given in Table 2.4.7 and E in Table 2.1.1. The discriminant function is

$$a' = [-.2547,\ -.1163,\ -2.1608,\ -.1431,\ .1376,\ .6685,\ .1992,\ -2.6915,\ -.1175,\ -.3834,\ .5770,\ .2301]$$

and

$$\underline{a}'\underline{z} = [-\ 2547\ z_1 - 1163\ z_2 - 21{,}608\ z_3 - 1431\ z_4 + 1376\ z_5 + 6685\ z_6 + 1992\ z_7 - 26{,}915\ z_8 - 1175\ z_9 - 3834\ z_{10} + 5770\ z_{11} + 2301\ z_{12}]\ .$$

2.5 Ordering

As expressed above we may regard the discriminant function as an artificial variable which is a linear combination of the observable variables. One may interpret the relative contribution of each observable variable to this artificial variable by computing the correlations of the artificial variable versus the observable ones. The correlation r_i of the artificial variable versus an observable one was evaluated using the relation (see Equation (1.2.8))

(2.5.1) $$r_i = (\underline{a}'\ E)_i / \sqrt{\underline{a}'\ E\ \underline{a}}\ \sqrt{e_{ii}}$$

Table 2.4.7

Matrix B All Variables

1	49.8840	1.2101	-51.0941
2	81.3360	4.1348	-85.4709
3	25.6413	2.4560	-28.0973
4	31.8943	2.7882	-34.6825
5	-57.6623	4.8951	52.7672
6	-31.9302	2.4607	29.4696
7	-21.7481	2.2249	19.5232
8	5.5677	.2918	- 5.8595
9	4.1024	- 6.1467	2.0443
10	15.4846	-18.4680	2.9834
11	4.5844	1.4907	- 6.0750
12	13.0248	-29.1103	16.0855

where $(\underline{a}' E)_i$ represents the i'th element of the row vector $\underline{a}' E$, and e_{ii} is the (i, i) diagonal element of the matrix E presented in Tables 2.1.1, 2.4.2, 2.4.4, or 2.4.6. Now in vector form $\underline{r}' = [r_1 \; r_2 \; \cdots \; r_p]$, hence

$$(2.5.2) \qquad \underline{r}' = \frac{1}{\sqrt{\underline{a}' E \underline{a}}} \; \underline{a}' E \; D_{1/\sqrt{e_{ii}}} \quad .$$

This $\underline{r}'$ was evaluated for each of the three sets and the set of all observable variables taken together. The results are as follows:

Set I (variables 1, 2, 3, 4, 11)

$$\underline{r}' = \begin{array}{ccccc} 1 & 2 & 3 & 4 & 11 \\ [.7669, & .7962, & .9022, & .7186, & .1478] \end{array}$$

Set II (variables 5, 6, 7, 8, 11)

$$\underline{r}' = \begin{array}{ccccc} 5 & 6 & 7 & 8 & 11 \\ [.4771, & .5065, & .4999, & -.2186, & -.1592] \end{array}$$

Set III (variables 9, 10, 12)

$$\underline{r}' = \begin{array}{ccc} 9 & 10 & 12 \\ [.3929, & .6069, & .2566] \end{array}$$

Set of all variables (1, 2, 3, ..., 12)

$$\underline{r}' = \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \\ [.5888, & .6109, & .6913, & .5507, & -.3327, & -.3533, \\ 7 & 8 & 9 & 10 & 11 & 12 \\ -.3483, & .1533, & .0246, & .0595, & .1126, & -.1295] \end{array} \quad .$$

The order will then be determined by the descending order of magnitude or modulus, thus

Set I : order - (3, 2, 1, 4, 11)

Set II : order - (6, 7, 5, 8, 11)

Set III : order - (10, 9, 12)

Set of all variables : order - (3, 2, 1, 4, 6, 7, 5, 8, 12, 11, 10, 9)

2.6 Selection

Clearly, from the construction of the variables, z_3 was expected to have the highest discriminating power (the standardized difference between two groups there was $(\mu_1^{(3)} - \mu_2^{(3)})/\sigma^{(3)} = (20 - 15)/\sqrt{2} \doteqdot 3.5$, and this value is larger than that for any other variable. On the other hand if, e.g., only three variables were to be selected for future purposes, the overall discriminant function would tell us to use 3, 2, 1 which as the factorial structure shows, are purely representative of one factor only (see discussion on anthropological and physiological measurements in the introduction). Making use of the fact that the twelve variables fall into three sets, our representative choice would be 3, 6, 10 . These are, in fact, the best discriminators in each set.

Chapter III: DEMONSTRATION STUDY 2

3.1 Description of Data

The data used in this study were taken from T. G. Thurstone's study given in [13]. The experimental units in this case were mentally retarded children from both public and special schools. These children were not institutionalized, but are in the lower range of the normal population. Many types of test were taken by these children and the findings are presented in [13]. We have chosen 12 of these tests as our observable variables. They are as follows:

z_1 = Binet Mental Age

z_2 = "Primary Mental Abilities Test" Mental Age

z_3 = Achievement Paragraph Meaning

z_4 = Achievement Word Meaning

z_5 = Achievement Spelling

z_6 = Achievement Arithmetic Reasoning

z_7 = Achievement Arithmetic Computation

z_8 = Gain Paragraph Meaning

z_9 = Gain Word Meaning

z_{10} = Gain Spelling

z_{11} = Gain Arithmetic Reasoning

z_{12} = Gain Arithmetic Computation .

Variables z_1 and z_2 are measured in months and variables $z_3, z_4, \ldots, z_{12}$ are measured in grade equivalents, i.e., a response of 2.0 would indicate performance at the second grade level. The children were placed in three age groups:

young (10 years 11 months and younger)
middle (11 years to 12 years 11 months)
old (13 years and older) .

Previous analysis of these data showed a significant interaction between school types and age groups; the older group performed much better in public schools than either the young or middle group. For this reason we have made three separate studies to find the discriminating variables between public and special schools; one study for each age group. The number of observations are shown in Table 3.1.1, a total of 480 experimental units.

Table 3.1.1

Number of Observations

	Young	Middle	Old	Total
Public	80	40	72	192
Special	120	60	108	288
Total	200	100	180	480

The matrix of corrected sums-of-squares and products pooled over the six groups is presented in Table 3.1.2 and is denoted as the matrix E . The matrix of correlations R is presented in Table 3.1.3. These matrices were obtained by the same methods as presented in Chapter II.

3.2 Factor Analysis

The Factor Analysis performed in this study was done in the same way as in Chapter II. An improved centroid solution for F was obtained and iterated into the maximum-likelihood estimate of F (25 iterations) with five factors. The final decision was five factors since α_5 was approximately .92 (normal curve equivalent of 1.41) indicating a good plausibility of five factors. The maximum-likelihood solution for F after 25 iterations is presented in Table 3.2.1.

Once F was obtained and the decision of the number of factors made, the "Principal Axes" (standard for comparison as indicated in Chapter I) was obtained by rotation, and is presented in Table 3.2.2; the communalities are also given there. The "Simple Structure" was then obtained by rotation for the purpose of identifying the variables belonging to common-factor sets. The Simple Structure is presented in Table 3.2.3. The five sets found were: A, which contains intelligence measures (variables 1, 2, 6, 7); B, which contains achievement measures (variables 3, 4, 5, 6, 7);

Table 3.1.2

Matrix E

	1	2	3	4
1	59450.00	49485.00	2088.80	2091.40
2	49485.00	92416.00	2944.20	2864.00
3	2088.80	2944.20	434.90	373.88
4	2091.40	2864.00	373.88	417.79
5	2275.10	3088.60	407.08	421.27
6	2952.40	3786.80	313.93	299.42
7	3229.30	4234.10	281.46	273.10
8	138.40	62.10	- 29.49	23.06
9	127.30	49.90	24.50	- 6.60
10	104.40	357.50	31.65	37.25
11	210.20	42.30	32.08	29.69
12	55.00	214.30	20.07	14.40

	5	6	7	8
1	2275.10	2952.40	3229.30	138.40
2	3088.6	3786.80	4234.10	62.10
3	407.08	313.93	281.46	- 29.49
4	421.27	299.42	273.10	23.06
5	665.19	367.64	350.28	40.34
6	367.64	414.96	373.12	24.75
7	350.28	373.12	554.53	16.91
8	40.34	24.75	16.91	141.25
9	42.30	36.40	33.42	41.48
10	22.90	31.07	29.20	23.53
11	47.50	- 5.70	39.91	18.53
12	22.09	21.60	16.51	8.95

Table 3.1.2 (Cont.)

	9	10	11	12
1	127.30	104.40	210.20	55.00
2	49.90	357.50	42.30	214.30
3	24.50	31.65	32.08	20.07
4	- 6.60	37.25	29.69	14.40
5	42.30	22.90	47.50	22.09
6	36.40	31.07	- 5.70	21.60
7	33.42	29.20	39.91	16.51
8	41.48	23.53	18.53	8.95
9	100.61	23.23	18.98	14.12
10	23.23	129.98	12.68	12.29
11	18.98	12.68	131.36	20.46
12	14.12	12.29	20.46	126.27

Table 3.1.3

Matrix of Correlations R

	1	2	3	4
1	1.000000	.667612	.410796	.419645
2	.667612	1.000000	.464407	.460915
3	.410796	.464407	1.000000	.877119
4	.419645	.460915	.877119	1.000000
5	.361786	.393927	.756854	.799114
6	.594424	.611499	.738984	.719116
7	.562432	.591459	.573138	.567388
8	.047760	.017188	- .118983	.094926
9	.052051	.016365	.117125	- .032192
10	.037557	.103149	.133119	.159849
11	.075219	.012140	.134217	.126736
12	.020074	.062733	.085645	.062695

	5	6	7	8
1	.361786	.594424	.562432	.047760
2	.393927	.611499	.591459	.017188
3	.756854	.738984	.573138	- .118983
4	.799114	.719116	.567388	.094926
5	1.000000	.699756	.576740	.131604
6	.699756	1.000000	.777827	.102230
7	.576740	.777827	1.000000	.060421
8	.131604	.102230	.060421	1.000000
9	.163511	.178147	.141489	.347956
10	.077880	.133783	.108763	.173656
11	.160690	- .024414	.147873	.136035
12	.076221	.094363	- .062393	.067016

Table 3.1.3 (Cont.)

	9	10	11	12
1	.052051	.037557	.075219	.020074
2	.016365	.103149	.012140	.062733
3	.117125	.133119	.134217	.085645
4	- .032192	.159849	.126736	.062695
5	.163511	.077880	.160690	.076221
6	.178147	.133783	- .024414	.094363
7	.141489	.108763	.147873	- .062393
8	.347956	.173656	.136035	.067016
9	1.000000	.203138	.165099	.125275
10	.203138	1.000000	.097040	.095932
11	.165099	.097040	1.000000	.158863
12	.125275	.095932	.158863	1.000000

Determinant of R = .001001523

Table 3.2.1

Matrix F (Maximum-Likelihood Solution, 25 Iterations)

	1	2	3	4	5
1	.628866	.230941	-.321293	.336145	.043374
2	.638647	.278587	-.265802	.326764	.015320
3	.809300	.284088	.431709	-.040941	-.258013
4	.809083	.321143	.420957	-.112223	.214127
5	.764734	.137118	.245682	-.186324	.054366
6	.870004	.319699	-.224076	-.278563	-.113919
7	.735920	.233001	-.229531	-.024823	-.041500
8	.248250	-.488106	-.166390	-.204546	.456236
9	.354375	-.696355	-.115519	-.123671	-.282028
10	.224312	-.179844	.039434	-.043212	.062182
11	.185520	-.265430	.180950	.199239	.042936
12	.120990	-.104362	.004798	-.063868	-.050182

Table 3.2.2

Principal Axes

	I	II	III	IV	V	h_i^2
1	.643	.079	-.460	.027	.188	.6669
2	.668	.129	-.408	.005	.185	.6631
3	.877	.131	.375	-.238	.085	.9903
4	.880	.106	.381	.240	.074	.9934
5	.787	-.081	.263	.073	-.035	.7017
6	.926	-.017	-.145	-.055	-.344	.9999
7	.763	-.010	-.239	-.019	-.102	.6509
8	.100	-.625	-.063	.414	-.036	.5775
9	.159	-.755	-.023	-.350	.012	.7187
10	.169	-.233	.055	.040	.048	.0899
11	.106	-.223	.099	-.017	.329	.1791
12	.092	-.137	.032	-.057	-.025	.0321

Table 3.2.3

Simple Structure

	A	B	C	D	E
1	.693	-.031	.017	.010	.035
2	.661	.026	-.025	-.008	-.027
3	.011	.758	.006	.065	-.330
4	-.004	.803	-.008	-.306	-.025
5	.023	.659	.134	-.069	.038
6	.372	.495	-.021	-.009	-.034
7	.458	.271	.030	.017	.017
8	.006	.003	.553	.021	.741
9	-.002	-.016	.751	.706	.375
10	-.005	.112	.247	.097	.193
11	.003	.035	.328	.181	.159
12	-.018	.068	.134	.112	.063

Transformation Vectors from F

I	II	III	IV	V
.361	.477	.347	.156	.176
.193	.248	-.935	-.653	-.652
-.709	.654	.015	-.124	-.277
.574	-.532	.053	.160	-.129
.023	.011	.052	-.712	.671

C, which contains gain measures (variables 8, 9, 10, 11); and the other two factors, D and E, of the Simple Structure are well overdetermined pseudo-factors showing a very high loading on a verbal gain variable and a negative one on the corresponding achievement variable. No such inverse relationship was observed in the arithmetic variables. These two factors, D and E, are probably due to the fact that the children in the lowest group on the achievement tests had scores so low in the first administration that they could not but gain in the second.

3.3 Discriminatory Analysis

The discriminant function between school types was sought for each of the three sets mentioned in Section 3.2, on each of the three age groups. Also, the overall discriminant function for all twelve variables on each of the three age groups was determined.

As presented in the previous chapters, the discriminant vector $\underline{a}$ is the eigenvector associated with the largest characteristic root of $E^{-1} H$, or

$$(3.3.1) \quad E^{-1} H \underline{a} = \lambda \underline{a} \quad .$$

Now, in this case,

$$(3.3.2) \quad H = \sum_{i=1}^{2} n_i (\overline{\underline{y}}_i - \overline{\overline{\underline{y}}})(\overline{\underline{y}}_i' - \overline{\overline{\underline{y}}}')$$

where $\overline{\underline{y}}_1$ is the mean vector for public schools, and $\overline{\underline{y}}_2$

the mean vector for special schools. This reduces to the form

$$(3.3.3) \quad H = \frac{n_1 n_2}{n_1 + n_2} (\underline{\bar{y}}_1 - \underline{\bar{y}}_2)(\underline{\bar{y}}_1' - \underline{\bar{y}}_2') ,$$

and hence

$$(3.3.4) \quad E^{-1} H = \frac{n_1 n_2}{n_1 + n_2} E^{-1} (\underline{\bar{y}}_1 - \underline{\bar{y}}_2)(\underline{\bar{y}}_1' - \underline{\bar{y}}_2') .$$

We shall denote $(\underline{\bar{y}}_1 - \underline{\bar{y}}_2)$ by $\underline{\bar{d}}$ and $(\underline{\bar{y}}_1' - \underline{\bar{y}}_2')$ by $\underline{\bar{d}}'$, hence

$$(3.3.5) \quad E^{-1} H = \frac{n_1 n_2}{n_1 + n_2} E^{-1} \underline{\bar{d}}\, \underline{\bar{d}}' .$$

Since $ch\ (A\ B) = ch\ (B\ A)$, the characteristic root of $E^{-1} H$ is the same as the characteristic root of $\frac{n_1 n_2}{n_1 + n_2} \underline{\bar{d}}' E^{-1} \underline{\bar{d}}$, which is a scaler quantity, hence

$$(3.3.6) \quad \lambda = \frac{n_1 n_2}{n_1 + n_2} \underline{\bar{d}}' E^{-1} \underline{\bar{d}}$$

the only non-zero characteristic root of $E^{-1} H$. Now by pre-multiplying Equation (3.3.6) by $E^{-1} \underline{\bar{d}}$ we get

$$(3.3.7) \quad \frac{n_1 n_2}{n_1 + n_2} E^{-1} \underline{\bar{d}}\, \underline{\bar{d}}' E^{-1} \underline{\bar{d}} = \lambda E^{-1} \underline{\bar{d}} ,$$

or

$$(3.3.8) \quad E^{-1} H (E^{-1} \underline{\bar{d}}) = \lambda (E^{-1} \underline{\bar{d}}) .$$

Hence,

$$(3.3.9) \quad \underline{a} = E^{-1} \underline{\bar{d}} .$$

This vector $\underline{a}$, which can, of course, be multiplied by any arbitrary constant, is equivalent to the discriminant function for two groups as given by R. A. Fisher (see [7]).

The matrix E for the young group is presented in Table 3.3.1, for the middle group in Table 3.3.2, and the old group in Table 3.3.3. The mean column vectors for both public and special schools on all three age groups are shown in Table 3.3.4, and the difference vectors, $\bar{\underline{d}}$, are shown in Table 3.3.5.

The discriminant functions $\underline{a}'\underline{z}$, are as follows: (variables z_1 and z_2 in months, variables $z_3, z_4, \ldots, z_{12}$ in grade (year) equivalents. The values of the $\underline{a}$'s were multiplied, in each case, by a convenient power of 10).

Young

Set I (variables 1, 2, 6, 7):

$$\underline{a}'\underline{z} = 3.32\ z_1 + .09\ z_2 - 50.80\ z_6 + 7.68\ z_7$$

Set II (variables 3, 4, 5, 6, 7):

$$\underline{a}'\underline{z} = 1.32\ z_3 - .31\ z_4 - .11\ z_5 - .15\ z_6 - .42\ z_7$$

Set III (variables 8, 9, 10, 11):

$$\underline{a}'\underline{z} = 1.50\ z_8 - .88\ z_9 - .90\ z_{10} - .67\ z_{11}$$

Middle

Set I :

$$\underline{a}'\underline{z} = .05\ z_1 - .54\ z_2 - 74.51\ z_6 + 46.52\ z_7$$

Table 3.3.1

Matrix E for Young (age: up to 10 yrs. 11 mo.)

	1	2	3	4
1	16929.00	17437.00	789.50	820.60
2	17437.00	31297.00	1210.90	1257.00
3	789.50	1210.90	127.30	117.09
4	820.60	1257.00	117.09	131.77
5	877.10	1417.40	136.21	140.10
6	837.10	1372.70	86.43	83.60
7	1006.10	1600.40	93.90	94.64
8	- 30.60	79.90	- 7.27	5.48
9	- 56.00	- 27.50	- .11	- 7.50
10	58.70	125.40	5.92	11.80
11	138.90	178.60	16.41	17.26
12	6.70	111.90	2.64	4.54

	5	6	7	8
1	877.10	837.10	1006.10	- 30.60
2	1417.40	1372.70	1600.40	79.90
3	136.21	86.43	93.90	- 7.27
4	140.10	83.60	94.64	5.48
5	217.45	114.30	119.21	11.59
6	114.30	106.07	99.61	11.63
7	119.21	99.61	148.88	9.22
8	11.59	11.63	9.92	42.73
9	8.45	8.59	6.46	21.89
10	- .33	6.33	9.90	10.09
11	20.70	4.21	24.11	5.48
12	8.14	5.69	- 8.76	6.11

Table 3.3.1 (Cont.)

	9	10	11	12
1	- 56.00	58.70	138.90	6.70
2	- 27.50	125.40	178.60	111.90
3	- .11	5.92	16.41	2.64
4	- 7.50	11.80	17.26	4.54
5	8.45	- .33	20.70	8.14
6	8.59	6.33	4.21	5.69
7	6.46	9.90	24.11	- 8.76
8	21.89	10.09	5.48	6.11
9	38.85	10.06	6.07	2.82
10	10.06	50.18	8.53	7.47
11	6.07	8.53	51.26	10.45
12	2.82	7.47	10.45	47.32

Table 3.3.2

Matrix E Middle (ages: 11 yr. to 12 yr. 11 mo.)

	1	2	3	4
1	11349.00	8922.00	331.40	258.90
2	8922.00	28755.00	678.30	590.90
3	331.40	678.30	87.57	75.59
4	258.90	590.90	75.59	76.93
5	421.00	648.10	89.60	85.11
6	548.20	831.90	70.13	60.12
7	584.50	974.80	67.20	63.73
8	- 3.90	- 71.10	- 7.08	- 1.29
9	134.40	69.70	5.15	1.38
10	47.10	202.50	11.11	11.51
11	45.90	12.90	10.44	9.80
12	38.30	11.00	2.80	.51

	5	6	7	8
1	421.00	548.20	584.50	- 3.90
2	648.10	831.90	974.80	- 71.10
3	89.60	70.13	67.20	- 7.08
4	85.11	60.12	63.73	- 1.29
5	147.26	82.37	85.50	5.41
6	82.37	95.74	84.63	- 2.38
7	85.50	84.63	148.19	- 6.59
8	5.41	- 2.38	- 6.59	23.75
9	12.56	12.05	10.80	9.32
10	14.39	16.70	14.14	3.49
11	14.55	- 1.02	6.47	4.22
12	5.74	9.03	- 1.31	1.58

Table 3.3.2 (Cont.)

	9	10	11	12
1	134.40	47.10	45.90	38.30
2	69.70	202.50	12.90	11.00
3	5.15	11.11	10.44	2.80
4	1.38	11.51	9.80	.51
5	12.56	14.39	14.55	5.74
6	12.05	16.70	- 1.02	9.03
7	10.80	14.14	6.47	- 1.31
8	9.32	3.49	4.22	1.58
9	17.83	1.60	3.63	3.27
10	1.60	29.96	- 1.18	2.62
11	3.63	- 1.18	30.12	2.79
12	3.27	2.62	2.79	20.06

Table 3.3.3

Matrix E Old (ages: 13 yr. up)

	1	2	3	4
1	31175.00	23126.00	967.90	1011.90
2	23126.00	32364.00	1055.00	1016.10
3	967.90	1055.00	220.03	181.20
4	1011.90	1016.10	181.20	209.09
5	977.00	1023.10	181.27	196.06
6	1567.10	1582.20	157.37	153.00
7	1638.70	1658.90	120.36	114.73
8	172.90	53.30	- 15.14	18.87
9	48.90	7.70	19.46	- .48
10	- 1.40	29.60	14.62	13.94
11	25.40	- 149.20	5.23	2.63
12	10.00	91.40	14.63	9.35

	5	6	7	8
1	977.00	1567.10	1638.70	172.90
2	1023.10	1582.20	1658.90	53.30
3	181.27	157.37	120.36	- 15.14
4	196.06	153.00	114.73	18.87
5	300.48	170.97	145.57	23.34
6	170.97	213.15	188.88	15.50
7	145.57	188.88	257.46	13.58
8	23.34	15.50	13.58	74.77
9	21.29	15.76	16.16	10.27
10	8.84	8.04	5.16	9.95
11	12.25	- 8.89	9.33	8.83
12	8.21	6.88	- 6.44	1.26

Table 3.3.3 (Cont.)

	9	10	11	12
1	48.90	- 1.40	25.40	10.00
2	7.70	29.60	-149.20	91.40
3	19.46	14.62	5.23	14.63
4	- .48	13.94	2.63	9.35
5	21.29	8.84	12.25	8.21
6	15.76	8.04	- 8.89	6.88
7	16.16	5.16	9.33	- 6.44
8	10.27	9.95	8.83	1.26
9	43.93	11.57	9.28	8.03
10	11.57	52.84	5.33	2.20
11	9.28	5.33	49.98	7.22
12	8.03	2.20	7.22	58.89

Table 3.3.4*

Means (Public Schools)

Var	Young	Middle	Old
1	76.9500	95.5000	110.0556
2	76.8250	97.0250	112.2847
3	1.8975	2.4425	3.3292
4	1.8688	2.4825	3.2569
5	1.8638	2.8200	3.6389
6	1.4850	2.4475	3.5139
7	1.7762	2.7450	3.8569
8	.4750	.3025	.2944
9	.4262	.3925	.2625
10	.5088	.3025	.2806
11	.4850	.3375	.2833
12	.5488	.4225	.3583

Means (Special Schools)

Var	Young	Middle	Old
1	79.2417	93.7000	106.2685
2	77.1417	93.8333	106.7685
3	1.2525	1.9050	2.5917
4	1.4108	1.9667	2.5768
5	1.4083	2.0367	2.7954
6	1.3125	2.0850	2.8426
7	1.7325	2.7533	3.4898
8	.5092	.3850	.3463
9	.4117	.3717	.3944
10	.4642	.5550	.3231
11	.4458	.4550	.4426
12	.5483	.3683	.3981

*Mental Age in months, achievement measures in grade equivalents.

Table 3.3.5

Mean Differences (Public minus Special)

Var	Young	Middle	Old
1	-2.2917	1.8000	3.7870
2	- .3167	3.1917	5.5162
3	.6450	.5375	.7375
4	.4579	.5158	.6801
5	.4554	.7833	.8435
6	.1725	.3625	.6713
7	.0438	- .0083	.3671
8	- .0322	- .0825	- .0518
9	.0146	.0208	- .1319
10	.0446	- .2525	- .0426
11	.0392	- .1175	- .1592
12	.0004	.0542	- .0398

Set II:

$$\underline{a}'\underline{z} = .72\ z_3 + 4.02\ z_4 + 5.13\ z_5 + 1.62\ z_6 - 6.00\ z_7$$

Set III:

$$\underline{a}'\underline{z} = 3.08\ z_8 - 4.52\ z_9 + 9.43\ z_{10} + 4.38\ z_{11}$$

<u>Old</u>

Set I:

$$\underline{a}'\underline{z} = .88\ z_1 - 1.07\ z_2 - 52.96\ z_6 + 25.89\ z_7$$

Set II:

$$\underline{a}'\underline{z} = 1.54\ z_3 - .70\ z_4 + 1.58\ z_5 + 3.25\ z_6 - 2.26\ z_7$$

Set III:

$$\underline{a}'\underline{z} = .04\ z_8 + 2.42\ z_9 - .01\ z_{10} + 2.73\ z_{11} \quad .$$

The discriminant functions for each age group on <u>all</u> variables are:

<u>Young</u>

$$\begin{aligned}\underline{a}'\underline{z} = {} & 3.68\ z_1 + .08\ z_2 - 181.68\ z_3 + 63.98\ z_4 + 18.47\ z_5 \\ & + .75\ z_6 + 34.58\ z_7 - 62.35\ z_8 + 41.11\ z_9 - 10.94\ z_{10} \\ & - 4.44\ z_{11} + 14.63\ z_{12}\end{aligned}$$

<u>Middle</u>

$$\begin{aligned}\underline{a}'\underline{z} = {} & .05\ z_1 + .11\ z_2 - 3.85\ z_3 + 10.48\ z_4 + 6.48\ z_5 \\ & + .12\ z_6 - 6.52\ z_7 - 5.80\ z_8 + 5.40\ z_9 - 13.54\ z_{10} \\ & - 8.47\ z_{11} + 3.01\ z_{12}\end{aligned}$$

Old

$$\underline{a}'\underline{z} = .36\ z_1 - .71\ z_2 - 40.78\ z_3 + 36.34\ z_4 - 25.71\ z_5 - 25.86\ z_6 + 20.87\ z_7 - 12.13\ z_8 + 58.15\ z_9 + 3.53\ z_{10} + 19.12\ z_{11} + 10.90\ z_{12} \quad .$$

3.4 Ordering

In order to make a representative selection we must first order the variables within each "common-factor" set as to their importance. This is done by ordering them according to their absolute correlation with the artificial variable $\underline{a}'\underline{z}$. These correlations are given by Equation (1.2.8), which states:

$$r' = \frac{1}{\sqrt{\underline{a}'\ E\ \underline{a}}}\ \underline{a}'\ E\ \ D_{1/\sqrt{e_{ii}}} \quad .$$

Now, since $\underline{a}'\ E = \underline{\bar{d}}'$,

$$(3.4.1) \qquad r' = \frac{1}{\sqrt{\underline{\bar{d}}'\ \underline{a}}}\ \underline{\bar{d}}'\ \ D_{1/\sqrt{e_{ii}}} \quad .$$

Note that r_i is in the form of a standardized mean. The correlations are then as follows:

Young

Set I:

$$\begin{array}{rcccc} & 1 & 2 & 6 & 7 \\ \underline{r}' = & [-.44, & -.04, & .42, & .09] \end{array}$$

Set II:

$$\begin{array}{cccccc} & 3 & 4 & 5 & 6 & 7 \\ \underline{r}' = & [\ .73, & .51, & .39, & .21, & .04\ ] \end{array}$$

Set III:

$$\begin{array}{ccccc} & 8 & 9 & 10 & 11 \\ \underline{r}' = & [\ -.44, & .21, & .56, & .48\ ] \end{array}$$

All variables:

$$\underline{r}' = [\ -.19,\ -.02,\ .62,\ .43,\ .34,\ .18,\ .04,\ -.05,\ .03,\ .07,\ .06,\ .00\ ]$$

<u>Middle</u>

Set I:

$$\begin{array}{ccccc} & 1 & 2 & 6 & 7 \\ \underline{r}' = & [\ .31, & .35, & .69, & -.01\ ] \end{array}$$

Set II:

$$\begin{array}{cccccc} & 3 & 4 & 5 & 6 & 7 \\ \underline{r}' = & [\ .68, & .70, & .76, & .44, & -.01\ ] \end{array}$$

Set III:

$$\begin{array}{ccccc} & 8 & 9 & 10 & 11 \\ \underline{r}' = & [\ -.30, & .87, & -.85, & -.38\ ] \end{array}$$

All variables:

$$\underline{r}' = [\ .14,\ .16,\ .48,\ .49,\ .54,\ .31,\ -.01,\ -.14,\ .00,\ -.41,\ -.18,\ .10\ ]$$

Old

Set I:

$$\begin{array}{cccccc} & & 1 & 2 & 6 & 7 \\ \underline{r}' & = & [\ .40, & .57, & .86, & .43\ ] \end{array}$$

Set II:

$$\begin{array}{ccccccc} & & 3 & 4 & 5 & 6 & 7 \\ \underline{r}' & = & [\ .86, & .81, & .84, & .80, & .40\ ] \end{array}$$

Set III:

$$\begin{array}{cccccc} & & 8 & 9 & 10 & 11 \\ \underline{r}' & = & [\ -.22, & -.72, & -.21, & -.82\ ] \end{array}$$

All variables:

$\underline{r}' = [\ .30,\ .43,\ .70,\ .66,\ .69,\ .65,\ .32,\ -.08,\ -.28,\ -.08,$
$-.32,\ -.07\ ]$

The orderings then:

Young

Set I : (1, 6, 7, 2)

Set II : (3, 4, 5, 6, 7)

Set III: (10, 11, 8, 9)

All : (3, 4, 5, 1, 6, 10, 11, 8, 7, 9, 2, 12)

Middle

Set I : (6, 2, 1, 7)

Set II : (5, 4, 3, 6, 7)

Set III: (9, 10, 11, 8)

All : (5, 4, 3, 10, 6, 11, 2, 1 or 8, 12, 7, 9)

<u>Old</u>

Set I : (6, 2, 7, 1)

Set II : (3, 5, 4, 6, 7)

Set III: (11, 9, 8, 10)

<u>All</u> : (3, 5, 4, 6, 2, 7 or 11, 1, 9, 8 or 10, 12) .

3.5 <u>Selection</u>

Now, if we wish to choose the three tests that are closest to the best discriminator (discriminant function) between schools, the overall discriminant function would indicate the three achievement tests: Paragraph Meaning, Word Meaning and Spelling (variables 3, 4, and 5) for all three age groups. It should be noted that these three variables all belong to the same common-factor; more specifically, they are all verbal achievement measures. While these three doubtlessly show the greatest relative difference between the public and special school groups, a future study based on these three alone would completely disregard the gain measures and intelligence measures originally considered.

A representative selection (one variable from each factor) would lead to the following choice:

<u>Young</u>; Binet Mental Age, Achievement Paragraph Meaning, Gain Spelling (variables 1, 3, 10)

<u>Middle</u>; Arithmetic Reasoning, Achievement Spelling, Gain Word Meaning (variables 6, 5, 9)

Old; Arithmetic Reasoning, Achievement Paragraph Meaning, Gain Arithmetic Reasoning (variables 6, 3, 11).

It must be noted that, for the middle and old group, Arithmetic Reasoning showed up as the strongest discriminator on the Intelligence factor (even though the test was denoted as an "Achievement" test, and had high loadings on the achievement factor, too). The achievement factors in this study are probably more strongly determined by the verbal tests (see Table 3.2.3). The Arithmetic Achievement tests were as clearly present on the Intelligence as on the Achievement Factor, on the former, they even represent the best discriminator.

If additional experiments were to be performed with a reduced set of variables, and if one were interested in those variables only which show the strongest relative difference between the two groups, the three verbal tests found by the overall discriminant analysis should be chosen. This subset would not, however, be representative of the whole set of original variables.

The choice of one of the representative sets (depending on age) would probably not lead to as strong a differentiation as the former, but all characteristics present in the original variables would certainly be represented in the reduced set.

BIBLIOGRAPHY

[1] Anderson, T. W. "An Introduction to Multivariate Statistical Analysis", Wiley, New York, 1958.

[2] Bargmann, R. E. "A Study of Independence in Multivariate Normal Analysis", Inst. of Stat., Univ. of North Carolina, No. 186, 1957.

[3] Bargmann, R. E. and Brown, R. H. "I. B. M. 650 Programs for Factor Analysis", Virginia Polytechnic Institute, July 1961.

[4] Bargmann, R. E. and Thigpen, C. "Lecture Notes on Methods of Multivariate Analysis", Virginia Polytechnic Institute, 1961 (being mimeographed).

[5] Brown, R. H. "A Comparison of the Maximum Determinant Solution in Factor Analysis with Various Approximate Solutions", M.S. Thesis, Virginia Polytechnic Institute, 1960.

[6] Dixon, W. J. and Massey, F. J. "Introduction to Statistical Analysis", McGraw-Hill, New York, 1957.

[7] Fisher, R. A. "The Use of Multiple Measurements in Taxonomic Problems", Ann. Eugen., Vol. 7, 1936.

[8] Heck, D. L. "Charts of Some Upper Percentage Points of the Distribution of the Largest Characteristic Root", Ann. Math. Stat., Vol. 31, 1960.

[9] Howe, W. G. "Some Contributions to Factor Analysis", Oak Ridge National Laboratories, ORNL 1919 (Physics), 1955.

[10] Posten, H. O. "Power of the Likelihood Ratio Test of the General Linear Hypothesis in Multivariate Analysis", Ph.D. Dissertation, Virginia Polytechnic Institute, 1960.

[11] Roy, S. N. "Some Aspects of Multivariate Analysis", Wiley, 1957.

[12] Thurstone, L. L. "Multiple Factor Analysis", Univ. of Chicago Press, 1940.

[13] Thurstone, T. G. "An Evaluation of Educating Mentally Handicapped Children in Special Classes and Regular Classes", Department of Health, Education and Welfare, Contract No. 168 (6452), 1960.

ACKNOWLEDGEMENTS

I wish to express my sincere appreciation to Dr. Rolf E. Bargmann, without whose interest, encouragement and advice, this thesis would not have been possible. Also I would like to thank Dr. Boyd Harshbarger and the faculty members of the Department of Statistics for their cooperation and encouragement, and especially Dr. Geoffrey A. Watterson for his advice and aid as the co-reader.

My thanks also go to   for excellent work in typing this paper.

I am indebted to   who supplied the data for the applied study in Chapter III.

Also, I would like to express my appreciation to the Department of Health, Education and Welfare for their financial support which made this report possible.

The vita has been removed from
the scanned document

ABSTRACT

The purpose of this thesis is a study of procedures of selecting variables in a multivariate experiment. The linear discriminant function is used as an artificial variable, its correlation on the observed variables is evaluated, and the absolute magnitude of these correlations decide the inclusion of a given variable in a subset.

These subsets are obtained by two different methods:

(a) the complete set of variables is subjected to a discriminant analysis, and the strongest correlates are chosen as the subset whose members are "closest" to the discriminant function,

(b) the set of variables is broken down into common-factor subsets, by factor analysis, and the strongest representative variates in each subset are selected as the "representative" set of variables, which are thus representative of all characteristics of the original variables. This type of "representative" selection is the proposal and it represents the major portion of the thesis.

Chapter I is a theoretical exposition containing the background and formulation needed. Chapter II presents an explicit demonstration study in which the structure is known. Computational details are explained and comparisons are made between the known structure and the structure obtained by the sampling

data. Chapter III represents the analysis of a study of data from an educational experiment on retarded children.