DOCUMENT RESUME

ED 021 791 24 SP 001 524

By- Schoenfeldt, Lyle F.

PROGRAM FOR TRAINING IN COMPUTER AND MULTIVARIATE APPLICATIONS TO EDUCATIONAL RESEARCH. FINAL REPORT

American Inst. for Research in Behavioral Sciences, Pittsburgh, Pa.
Spons Agency- Office of Education (DHEW), Washington, D.C. Bureau of Research.
Bureau No- BR-6-2084
Pub Date Jun 67
Grant- OEG-1-6-062084-1789
Note- 37p.
EDRS Price MF-$0.25 HC-$1.56
Descriptors- *COLLEGE FACULTY, *COMPUTER SCIENCE EDUCATION, *EDUCATIONAL RESEARCH, *INSTITUTES (TRAINING PROGRAMS), POST DOCTORAL EDUCATION, PROGRAM CONTENT, *RESEARCH METHODOLOGY, STATISTICAL ANALYSIS
Identifiers- Project TALENT

Three postdoctoral fellows completed a 38-week training program designed to familiarize scientists already experienced in educational research with the techniques of designing and executing a large-scale, long-range educational research project. The program was conducted by the research staff of Project TALENT, a project of the Institute for Research in Education of the American Institutes for Research. Trainees participated in a series of 4 seminars: Project TALENT Seminar, Computer Applications to Educational Research, Statistical Analysis, and Research Methodology Applicable to Large-Scale Educational Research. In addition, each trainee conducted an individual research effort. Among the factors contributing to the success of the program were the interaction between participants and the research community in general, the individualization of the program, and the computer facilities available for trainee use. All 3 postdoctoral fellows have received faculty appointments at institutions of higher education and thus will have an opportunity to contribute to the training of other research workers. Appended are abstracts of the research accomplished by 2 of the trainees and 3 joint papers produced by 2 of them. (JS)

FINAL REPORT
Project No. 6-2084
Grant No. OEG 1-6-062084-1789

PA 24

PROGRAM FOR TRAINING IN COMPUTER AND MULTIVARIATE APPLICATIONS TO EDUCATIONAL RESEARCH

June 1967

U. S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
Bureau of Research

PROGRAM FOR TRAINING IN COMPUTER AND MULTIVARIATE APPLICATIONS TO EDUCATIONAL RESEARCH

Project No. 6-2084
Grant No. OEG 1-6-062084-1789

Lyle F. Schoenfeldt

June 1967

The research reported herein was performed pursuant to a grant with the Office of Education, U. S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

American Institutes for Research

Pittsburgh, Pennsylvania

INTRODUCTION

This report discusses the operation of an academic year (38 weeks) postdoctoral training program. The program was initiated September 1, 1966 and terminated May 31, 1967. Three postdoctoral fellows were selected for and completed the program. The objectives of the program were to familiarize scientists already experienced in educational research with the techniques of designing and executing a large-scale, long-range educational research project. The specific competencies developed by the program were as follows:

1. an understanding of computer techniques and capabilities;
2. statistical procedures applicable to large-scale educational research; and
3. research strategy appropriate with large data files.

The training program was conducted in the research setting of Project TALENT, a project of the Institute for Research in Education of the American Institutes for Research.

Description of the Program

The program was conducted by the research staff of Project TALENT. The training program consisted of a series of four seminars, in which each trainee participated. In addition, an individual research effort has been, or is in the process of being, completed by each of the three postdoctoral fellows. This project used a portion of the data collected in conjunction with Project TALENT.

The four seminars are described below.

1. Project TALENT Research Seminar, Chairman: Marion F. Shaycoft.

   This seminar included 1) background information about Project TALENT; 2) discussion of the sampling procedure and the sample; 3) the tests, inventories, and questionnaires used in conjunction with Project TALENT: 4) discussion of findings from past research, using Project TALENT data, and a discussion of current research. In the area of past and present research, findings concerning the American high school student and the American high school were presented; also findings based on the follow-ups one year and five years after graduation from high school. Problems in psychometric theory were discussed with special reference to the manner in which they impinged on Project TALENT research and the solutions that have been applied.

2. Seminar on Computer Applications to Educational Research, Chairman: Paul R. Lohnes.

   This seminar included the following topics: 1) programming considerations involved in generating correlation matrices, inverting symmetric matrices, and finding their eigenvalues and eigenvectors. Each of the three participants became conversant with the computer language FORTRAN, the problems involved, and what the operating system does in compiling and excuting a program; 2) the details of a large-scale computer installation with associated features; and 3) technical considerations in organizing, maintaining, updating, and effectively using a large-scale data file.

3. Seminar on Statistical Analysis, Chairman: Charles E. Hall. The following topics were presented to the three trainees in approximately the given order: 1) correlational analysis; 2) principal components analysis, principal factor analysis, mechanized rotational procedures; 3) multiple and canonical correlation; 4) central limits theorem and the variance ratio; 5) student's t-test and simple factorial univariate analysis of variance; 6) the general linear hypothesis model, and 7) multivariate analysis of variance with discriminant analysis as a subtopic.

4. Seminar on Research Methodology Applicable to Large-Scale Educational Research, Chairman: William W. Cooley. In this seminar Project TALENT scientists and the postdoctoral trainees discussed the methodological considerations involved in ongoing research. Successive sessions of this seminar were devoted to presentation of research being conducted by various members of the Project TALENT staff. In addition, each of the postdoctoral participants presented plans and progress regarding their own research with the Project TALENT data.

Evaluation of the Program

In general, all aspects of the training program were undertaken and accomplished as originally planned. The objectives were found to be quite realistic for a nine-month training effort. The fact that there were two times as many research staff members directly involved in the training program as there were trainees participating resulted in both a comprehensive and an individualized program of

instruction. At the time the program was proposed, it was realized that such an undertaking, no matter how ambitious, could not hope to fulfill the needs of the educational community for persons skilled in the computer and multivariate applications to educational research. For this reason, one of the selection factors was potential for contribution to the training of other research workers. Each of the three post-doctoral fellows selected for participation in the program has received faculty appointments at institutions of higher education and thus will have an outstanding opportunity to contribute to the training of other research workers.

Several features of the program deserve special mention. First, is the support provided by the TALENT staff with regard to the individual research undertaken by each of the three participants. In an effort to facilitate this research, the services of the Editorial Assistant, the Research Assistants, Computer Programmers, and many others were made available to the postdoctoral fellows. Another feature worthy of mention is the facilities that were made available to the fellows. Each was provided with virtually unlimited access to the several computers regularly utilized by the staff of Project TALENT.

An unanticipated, but nevertheless welcome, feature was the opportunity for the three postdoctoral fellows to interact with the research community in general. An example of this was the opportunity the three postdoctoral fellows had to spend an afternoon with Bert Green, Chairman of the Department of Psychology, Carnegie-Mellon University. In that afternoon, they were briefed on the

advanced work underway at the Carnegie-Mellon University with regard to the application of computers to behavioral research. In addition, each of the three fellows was provided opportunities to interact with the faculty of many departments of the University of Pittsburgh. Among the departments making faculty members available for discussion with the postdoctoral fellows were the Department of Educational Research, the Computer Center, the Department of Sociology, the Political Science Department, the Knowledge Availability Center, the Learning and Research Development Corporation (a R&D Center established by the OE and directed by Robert Glaser), and the Business School. Interaction with members of the research community provided a special opportunity for the postdoctoral fellows to put into perspective the individual research they undertook.

Still another feature worthy of mention was the individualization of the program. Aside from the four ongoing seminars, each of the three postdoctoral fellows had ample opportunity to work with those research staff members with interests similar to theirs, or capabilities uniquely associated to their individual research. The fact that there were six research staff members and three postdoctoral fellows enabled the instruction to be done at a much more individual and personal level than would have otherwise been possible. One last strength of the program deserving mention is the quality of the three postdoctoral participants. Whereas the late announcement of the initial awards handicapped other programs in selecting students, it was not an especially potent factor in effecting the quality of this program. An immediate and hardhitting publicity campaign following the announcement of support for the program produced widespread

interest and numerous applications for participation. As a result, it was possible to select from the applicants the three candidates who best met the criteria established in the proposal: 1) unusual career achievements; 2) the ability to benefit from the proposed training program; and 3) interest in, and potentials for outstanding contributions, to educational research and to the training of other research workers.

The major difficulty encountered in the program was the speed with which seminar and research activity had to proceed to provide indepth coverage of the material presented. Ideally, the program would have been of slightly longer duration to provide the opportunity for the postdoctoral fellows to more thoroughly assimilate the topics covered.

The overall evaluation of the program is highly favorable. Objective evidence to support this evaluation is from three sources. First are the products of the three postdoctoral fellows. Attachment A of this report includes an abstract of the research accomplished by two of the postdoctoral trainees. Attachment B includes three joint papers produced by two of the postdoctoral fellows as a direct result of their participation in the program. A second source of objective evidence is the positions obtained by the three fellows upon completion of their postdoctoral training. As mentioned, all three have joined the faculty of institutions of higher education and, thus, will have many and continued opportunities to embellish and disseminate the experiences garnered during the course of their postdoctoral education. The third source of evidence is the opinions

of the three postdoctoral fellows. Each was provided with several opportunities to evaluate the progress of the program during the course of the nine months. Suggestions made for improvements were incorporated into the program whenever possible. At the conclusion of the program the trainees were informally asked to give their opinions of the overall program. All three were quite positive in their evaluation of the experience gained in the course of the post-doctoral program. The major criticism concerned the short duration of the program.

The biggest disappointment on the part of those concerned with this program is the fact that it will not be permitted to continue. The original proposal outlined 4 one-year postdoctoral programs, the last three of which would have built on the experience gained from the first. We feel that we have both put together a good program and acquired the experience necessary to expand it. Despite this, we have been assigned no postdoctoral fellows for the coming academic year.

It should also be mentioned that during the course of the past nine or ten months we have had serious inquiries regarding our program from approximately 30 persons. In addition to these persons, there were many qualified applicants who, because of the short notice, were unable to apply for participation during the past year. In light of the success in both enrolling three postdoctoral fellows and offering them a well-planned nine-month program, the current procedures incorporated by the Research Training Branch of the U.S. Office of

Education, make little sense. The necessity for curtailing the postdoctoral aspect of the research training program is understandable. The reason that the Project TALENT program will not be allowed to continue is hard to understand. If the Office of Education continues to select postdoctoral fellows by means of national competition it is suggested that efforts be made to provide qualified institutions with a greater opportunity of acquiring fellows interested in being located at that institution.

Program Reports

1. Publicity

In addition to the announcement published in the AERA's Educational Researcher, the announcement included as Attachment C was sent to approximately 1200 persons from the Project TALENT mailing list in late June, 1966. The 1200 persons included the Project TALENT regional coordinators, college professors, and other professionals who have, from time to time, indicated interest in Project TALENT.

2. Application Summary

a. Approximate number of inquiries from prospective trainees: 15.

b. Number of completed applications received: 8.

c. Number of first-rank applications: 5.

d. How many applicants were offered admission: 4.

3. Trainee Summary

a. Number of trainees initially accepted to the program: 3.

Number of trainees enrolled at beginning of program: 3.

Number of trainees who completed program: 3.

b. Categorization of trainees

Number of trainees who are principally elementary or secondary public school teachers: 0

Number of trainees who are principally local public school administrators or supervisors: 0

Number of trainees from colleges or universities, junior colleges, research bureaus, etc.: 3

4. Program Director's Attendance

As described earlier, the program covered a nine-month period beginning September 1, 1966 and concluded May 31, 1967. The trainees were present continuously during this nine-month interval, except for the normal holiday and vacation schedule applicable to employees at the American Institutes for Research. The Director and all research staff of Project TALENT were present in accordance with outlined policy.

5. Financial Summary

	Budgeted	Expended or Committed
a. Trainee Support		
(1) Stipends	$ 8,500/per trainee	$25,500
(2) Dependency Allowance	0	0
(3) Travel (Relocation)	500/per trainee	1,500
b. Direct Costs (Institutional Allowance)	3,000	3,000
c. Indirect Costs	0	0
TOTAL	$30,000	$30,000

Attachment A

Abstracts of Projects of Research Fellows

Effect of Negro Density on Student Variables and the Post-High School Adjustment of Male Negroes

David E. Kapel

The major concern of this study was to evaluate the effects of Negro density, community, and regional differences on post-high school adjustment and student factors for Negro males. Three specific null hypotheses were tested. Two were rejected as a result of analyses that found: (1) environmental-parameter groups could be distinguished from each other; and (2) significant differences were generated by regional influences, but not by community and Negro density factors. The third null hypothesis was not rejected as a result of the analyses that found no significant environmental factors influencing types of post-high-school education acquired and projected.

The rejection of the first two hypotheses might have been a function of the mediating influence of environmental factors on student and employment variables, vis-a-vis social status, amounts spent on education, quality of education, and occupational opportunities across environmental levels; while the nonrejection of the third hypothesis indicated that environmental factors did not significantly influence the educational goals that were studied. It is also apparent that certain variables provided better discriminatory power than others, and that a multivariate approach gives a clear picture of the important and significant variables that need to be studied.

Role Expectancies for American Adolescents

William A. Love, Jr.

This study deals with the relationship between personality abilities, sex and sociometric standing. The researcher attempted to define role expectancies for American adolescents. Since sociometric status may be taken as an index of the acceptance accorded an individual, then if personality and ability traits held by these persons are analyzed, those traits which are valued can be assessed. Since this study considered both same sex and cross sex choices, the researchers were able to get some idea of what was valued within sex and cross sex.

The second aspect of the study was methodological. Techniques utilizing canonical correlation, which were developed by Douglas K. Stewart and this researcher were utilized in the analysis. Since these techniques are new, this study functioned as a try-out for their usefulness.

Attachment B

Joint Papers by Research Fellows

ASSESSING THE RELATIVE IMPORTANCE OF VARIABLES IN THE CANONICAL SOLUTION[1]

William Love and Douglas Stewart

Canonical correlation has proved worthwhile in various studies of behavioral data. Because a canonical correlation is the correlation between two linear composites, the correlation does not inform us of the relative importance of individual variables. The interpretation of a given canonical correlation is greatly aided by following Meredith's (1964) suggestion that the correlation between an observed variable and the canonical variate be computed (hereafter referred to as a "canonical loading"). Consider two sets of variables designated P and Q (for convenience the P set will be considered the predictor set and the Q set will be considered the criterion). Given a variables by canonical variates matrix of squared loadings (L), L_{ij} represents the proportion of variance of the i*th* variable associated with the j*th* canonical variate. Noting that $r^2_{ik}=r^2_{ij}\cdot r^2_{jk}$ (where $r^2_{ik\cdot j}=0$) we may multiply the squared canonical loading (L_{ij}) by the squared canonical correlation (λ_j) in order to determine the proportion of variance of the i*th* variable of the Q set predicted from the j*th* canonical composite of the P set. If for the i*th* variable we sum the proportions predicted from each of the canonical composites of the P set, we have the total proportion

[1]The authors wish to express their appreciation to Paul R. Lohnes who encouraged and guided the present effort while they were Office of Education Post-Doctoral Fellows (O.E.G. 1-6-062084) at Project TALENT.

of variance in the ith variable predicted by the canonical solution. Thus, if all canonical roots are extracted, this sum is the value of the squared multiple correlation between the ith variate of the Q set and all the variables of the P set.

Where L_q is a matrix of squared canonical loadings of M_q variables, λ is a column vector of squared canonical correlations, and H_q is a column vector of squared multiple correlations in the case of the full canonical solution (i.e., all canonical roots removed):

$$H_q = L_q \lambda$$

The mean of the elements of H_q can be interpreted as the proportion of variance in the Q set predicted from the P set (designated $\bar{R}$). It will also be noted that the column sum of squared loadings for the ith column when multiplied by the ith squared canonical correlation and divided by M_q (the rank of the Q set) is interpretable as the proportion of variance in the Q set predicted by the ith canonical root from the P set, and is therefore instructive in determining which canonical roots bear interpretation.

To demonstrate the techniques described above, the authors have reanalyzed data presented by Lohnes (1966), who factored two sets of measures which he termed: 1. Abilities (designated L) and 2. Motives (designated R).

The factors of the abilities domain are: 1. Verbal Knowledge; 2. Perceptual Speed and Accuracy; 3. Mathematics; 4. Hunting-Fishing; 5. English Language; 6. Visual Reasoning; 7. Color, Foods; 8. Etiquette; 9. Memory; 10. Screening; 11. Games. In the motives

domain: 1. Business Interests; 2. Conformity Needs; 3. Scholasticism; 4. Outdoors, Shop Interests; 5. Cultural Interests; 6. Activity Level; 7. Impulsion; 8. Science Interests; 9. Sociability; 10. Leadership; 11. Introspection.

Table 1 shows the canonical loadings and correlations for the two sets. Given that $M_L=M_R$ where M is the rank of the sets, all variance is extracted from both sides. Table 2 presents the column vectors H_L and H_R which contain squared multiple correlations. The mean of the first column ($\bar{R}$) is interpretable as the proportion of set variance predicted by the variables of the opposing set. Column 2 presents each squared multiple correlation as a proportion of the sum of the first column and therefore can be interpreted as the proportion of $\bar{R}$ attributable to each variable. The proportion of left variance predicted by the right set of variables ($\bar{R}_{L.R}$) and the proportion of right variance predicted by the left set of variables ($\bar{R}_{R.L}$) are both approximately .10, indicating relative independence between the two sets. The proportioned R^2 (column 2 of Table 2) for each variable is useful for describing the area of redundant variance. In the abilities (left) set, Verbal Knowledge (.270), Mathematics (.207), and English Language (.121) are the important variables. In the motives (right) set, Scholasticism (.241) and Science Interest (.152) are the major contributors. While the overlap between the two systems is approximately 10 per cent, the area of overlap tends to be the result of the relationship between academic ability variables in the left set, and academic interest variables in the right set.

The problem to which this paper has been addressed is the assessment of the relative importance of various variables in the canonical solution. We have suggested a summary measure for determining the proportion of variance of one set predicted by another set ($\bar{R}$). The relative contributions of variables to the general index have therefore been proposed as an indication of the relative importance of the variables to the canonical solution. It should be emphasized that $\bar{R}$ is the mean of squared multiple correlations only when all roots are removed (which is to say H_q contains R^2 s when all roots are considered but is smaller if fewer than M_q roots are considered). Researchers may on occasion wish to impose criteria as to which roots are used (such as significance levels) such that $\bar{R}$ is no longer the mean of squared multiple correlations.

REFERENCES

Lohnes, P. "Measuring Adolescent Personality." Pittsburgh: Project TALENT, 1966.

Meredith, W. "Canonical Correlations with Fallible Data." Psychometrika, XXIX (1964), 55-65.

Table 1

Canonical Loadings and Correlations

Left Set - Columns Are Canonical Factors, Rows Are Tests

1	.766	-.127	.067	-.364	-.349	.327	-.016	-.026	-.129	-.119	.011
2	.117	-.180	.105	.516	-.483	.076	-.083	.010	.463	.386	.259
3	.546	.455	-.085	.556	.411	.007	-.024	.006	-.080	.014	-.051
4	-.113	.242	-.480	-.029	-.330	-.114	-.689	.074	-.255	.156	-.084
5	.304	-.567	-.068	-.079	.316	-.612	-.142	-.080	-.033	.248	.107
6	.075	.142	-.595	-.235	.062	-.074	.287	.437	.389	- 048	.358
7	.076	.015	.041	.045	-.234	-.146	.355	.533	-.176	.390	-.567
8	-.046	-.321	.016	.368	-.137	-.072	-.020	.444	-.392	-.582	.213
9	.087	-.027	-.135	.049	-.095	-.232	-.109	-.123	.536	-.511	-.577
10	-.066	-.198	-.562	.249	-.164	.108	.450	-.498	-.272	-.007	-.120
11	-.072	-.447	-.155	.065	.401	.614	-.239	.257	.127	.116	-.275
R_c	.664	.445	.393	.336	.302	.201	.132	.126	.054	.036	.010

Right Set - Columns Are Canonical Factors, Rows Are Tests

1	-.169	-.189	.334	.401	.206	-.337	-.398	-.261	.301	.004	.436
2	.236	-.473	-.268	.379	-.432	-.091	.491	-.182	.105	.054	.147
3	.763	.188	.136	.259	.141	-.320	.007	.340	-.066	.190	-.137
4	-.116	.302	-.564	-.045	-.350	-.541	-.297	.043	-.002	-.256	.064
5	.262	.002	.459	-.585	-.338	-.162	.135	.076	.234	-.186	.353
6	-.337	.365	.081	.293	-.203	.166	.248	.547	.444	.142	.118
7	.029	.356	.362	.285	-.552	.170	-.158	-.214	-.501	.061	.028
8	.552	.291	-.288	.096	.125	.501	-.215	- 093	.194	-.190	.352
9	-.058	-.555	-.019	-.013	-.318	.353	-.506	.436	-.021	-.131	.024
10	.141	-.002	.231	.290	-.003	-.042	.035	.006	.202	-.714	-.537
11	.048	.078	-.034	-.191	-.285	.139	-.204	-.419	.561	.386	-.432

Table 2

Left

Variable	R^2	$R^2_i / \Sigma R^2_i$
1	.293	.2[illegible]0
2	.067	.062
3	.224	.207
4	.072	.066
5	.131	.121
6	.073	.067
7	.016	.015
8	.043	.039
9	.011	.010
10	.076	.070
11	.078	.072

2

$$\frac{\Sigma R^2}{M} = \bar{R}_{L.R} = .098$$

Right

Variable	R^2	$R^2_i / \Sigma R^2_i$
1	.068	.058
2	.118	.101
3	.282	.241
4	.098	.084
5	.114	.097
6	.098	.084
7	.086	.073
8	.177	.152
9	.084	.072
10	.027	.023
11	.018	.015

$$\frac{\Sigma R^2}{M} = \bar{R}_{R.L} = .106$$

A SIMPLE ALGORITHM FOR COMPUTING MULTIPLE CORRELATIONS FROM THE CANONICAL SOLUTION[1]

William Love and Douglas Stewart

Canonical correlations are used increasingly by behavioral researchers. Following Meredith (1964) many analysts choose to interpret the correlations between observed variables and canonical variates (hereafter referred to as canonical loadings) rather than the weights which form the canonical variates. Given two sets of variables (designated p and q), the multiple correlations between each element of one set and all elements of the opposing set can be simply computed. Given a matrix of squared canonical loadings (L_p, where L_p is a variable by canonical variate matrix for the p set) and a column vector of squared canonical correlations (λ),

$$R_p = L_p\lambda$$

where R_p is a column vector of squared multiple correlations between each element of the p set and all elements of the q set. Thus, in order to compute squared multiple correlations:

1. Square each element of a canonical loading matrix (forming L_p);
2. Multiply each element of the jth column of L_p by the square of the jth canonical correlation (λ_j);
3. The sum of the elements in the ith row is the squared multiple correlation of the ith variable of the p set with the variables of the q set.

It has also been noted that the sum of the jth column of this matrix when divided by M_p (the number of variables in the p set) can be interpreted as the proportion of variance in the p set accounted for by the jth canonical root and is therefore instructive in determining which canonical roots bear interpretation (two linear composites may be well correlated without representing significant portions of variance).

---

[1]This work was undertaken while the authors were Office of Education Post Doctoral Fellows (O.E.G 1-6-062084) at Project TALENT.

A GENERAL CANONICAL CORRELATION INDEX

Douglas Stewart and William Love

Because a canonical correlation is the correlation between two linear composites, it presents some interpretive problems. No measure of the redundancy in one set of variables, given another set of variables, has been available. A nonsymmetric index of redundancy is proposed which represents the amount of predicted variance in a set of variables.

A GENERAL CANONICAL CORRELATION INDEX[1]

The interpretation of canonical correlations presents some problems. Whereas a squared multiple correlation represents the proportion of criterion variance predicted by the optimal linear combination of predictors, a squared canonical correlation represents the variance shared by linear composites of two sets of variables, and not the shared variance of the two sets.

Unfortunately, therefore, canonical correlations cannot be interpreted as correlations between sets of variables. It is important to note that a relatively strong canonical correlation may obtain between two linear functions, even though these linear functions may not extract significant portions of variance from their respective batteries. This is the problem of interpretation to which this paper is addressed.

Rozeboom (1965) has suggested the relevance of information theoretic concepts in dealing with canonical correlations. Uncertainty and alienation are considered parallel, and similarly, redundancy and correlation are treated as analogous. Given this approach, Rozeboom develops a general index which is similar to one presented by Anderson (1958, p. 244). Both measures are symmetric, i.e., given two sets of variables, one number is presented which presents the magnitude of their intersection. A directional or non-symmetric index is possible by pursuing the information theoretic analogues suggested by Rozeboom. In addition to the primitive concept of uncertainty (or entropy) Shannon (Shannon and Weaver, 1949) discusses conditional uncertainty.

---

[1]The authors wish to express their appreciation to Paul R. Lohnes who encouraged and guided the present effort while they were Office of Education Post-Doctoral Fellows (O.E.G. 1-6-062084) at Project TALENT

Similarly, one may discuss the complement of conditional uncertainty as conditional redundancy. A non-symmetric measure is considered desirable because one set of variables may be almost completely sub-subsumed by a larger set; i.e., redundancy can be represented as the intersection of two sets of variables, and it is desirable to represent the proportion of one set which is in the intersection (see Fig. 1).

INSERT FIGURE 1

In the case pictured in Figure 1, it is clear that most of set A is contained in set B, whereas a relatively large portion of set B is outside the intersection. This paper proposes an index based on canonical correlation which is non-symmetric and has been worthwhile in the analysis of various partitioned matrices.

If we were to factor analyze two sets of variables independently and then develop weights which would rotate the two factor structures to maximum correlation, we would have a canonical solution (Hotelling, 1935). In the canonical case the factors are usually referred to as canonical variates. The correlation between the first factor of the left set and the first factor of the right set is the first canonical correlation $\left(R_{c_1}\right)$. In order to take advantage of the well developed language of factor analysis, we shall call them canonical factors.

Since the complete factor structure of a set of variables will contain as many factors as there are variables,[1] it is obvious that if

---

[1]This is only true where the rank of the matrix equals the order. In general this is the case and will be assumed in this paper.

the larger set is composed of five variables and the smaller set of three variables, only three factors can be extracted from the smaller set. As a result, R_c's are available between three of the factors of the larger set and the three factors of the smaller set. The remaining two factors in the larger set have no counterpart in the smaller set and do not enter into the canonical solution.

In the traditional interpretation of canonical correlations, the magnitude of the R_c's, whether or not they are significantly non-zero, and the weights used to obtain the R_c's are considered (Cooley and Lohnes, 1962). The interpretation of these weights has all the problems attendant to the beta weights of common multiple regression. At the suggestion of Meredith (1964), some investigators now compute the correlations between the variables in a set and the canonical factors of that set (the factor loadings of factor analytic parlance).[1]

Before we consider a method of calculating an index of redundancy we should agree on vocabulary. We need one index for the redundancy in the left set given the right and another index for the reverse relation. For the sake of simplicity, we will consider one set of variables as the predictor or conditioning set and the other set as the criterion, as in multiple regression. We talk about the proportion of variance in the criterion accounted for by the predictors, but seldom if ever consider the reverse relationship. It is obvious that by reversing our definition of criterion and predictor we could develop the index going in the other direction. The canonical factors

---

[1]This proposal will be utilized in the forthcoming second edition of Cooley and Lohnes.

of the predictor set will be FP_i and similarly FC_i for the criterion set. The variables of the predictor and criterion sets will be P_i and C_i, respectively. Since the index about to be proposed utilizes the concept of a factor extracting a proportion of the variance (more appropriately proportion of trace) of a set of variables (usually a battery of tests), we will define the column sum of the squared loadings of variables within a set on a canonical factor of the set as the variance extracted by that factor. When this is divided by the number of variables in the set (M), the resulting value is the proportion of the variance of the set extracted by that canonical factor. This will be symbolized as VP_i and VC_i. The squared canonical correlations $\left(R_{c_i}^2\right)$ will be written as λ_i (following Cooley and Lohnes, 1962). This is the proportion of variance in one of the ith pair of canonical variates predictable from the other member of the pair. If the VC_i is multiplied by the λ_i, the resulting figure is the proportion of the variance of the C set explained by correlation between FP_i and FC_i. If this value is calculated for each of the M_c pairs of canonical factors, the result is an index of the proportion of variance of C predictable from P, or the redundancy in C given P.

$$\overline{R} = \sum_{k=1}^{M_c} \lambda_k \, VC_k = \sum_{k=1}^{M_c} \lambda_i \left[\sum_{j=1}^{M_c} \left(L_{jk}^2 \Big/ M_c \right) \right]$$

(where L_{jk} is the correlation between the jth variable and kth canonical factor.)

We have called this index $\bar{R}$ (R bar) because it was noted that it a mult R^2 were computed between the total P set and each variable of the C set, $\bar{R} = \Sigma R^2/Mc$. In other words $\bar{R}$ is the mean squared multiple correlation. The possible range of $\bar{R}$ is from 0.0 to +1.[1]

An example of the use of canonical correlation is presented by Lohnes and Marshall (1965).[2] In this study three scores from the Pintner General Ability Test (PGAT) and ten from the Metropolitan Achievment Test were entered into a canonical correlation with the 7th and 8th year course grades in English, arithmetic, social studies, and science of 230 junior high school students in a small, rural college town. The first two canonical correlations were reported (R_{c_1} = .90 and R_{c_2} = .66). The canonical weights were reported and interpreted.

In the present analysis of the Lohnes-Marshall data, the weights were ignored and the factor loadings and $\bar{R}$'s were inspected.

INSERT TABLE 1

In the left set, loadings from .707 to .917 are found on the first factor. The loadings on the second factor drop substantially. The same condition holds in the right set. In Table 2, columns 1 and 2 present the canonical correlations and their squares. Note that the

---

[1]It should be noted that if $Mc \leq Mp$ then $\bar{R} \leq 1.0$. If R is calculated for P and $Mc < Mp$ then $\bar{R} < 1.00$. The only time $\bar{R}$ can equal 1.0 is when each $\lambda = 1.00$ and the canonical factors of the set in question extract 100 percent of the generalized variance in that set.

[2]Professor Paul R. Lohnes graciously allowed us to use his data and modified his latest canonical program to calculate our index.

upper portion of Table 2 considers the left set as criterion and right set as the predictor set, while the lower portion reverses these roles. The third column of Table 2 presents the proportions of the variance of the set extracted by each canonical factor (variate). The fourth column is the amount of redundant variance attributed to each canonical factor. The fifth column expresses the values in the fourth column as proportions of the total redundancy.

From this we see that:

1. The eight canonical factors extract 90 percent of variance of the left set;
2. Fifty-nine percent of the variance of the left set is predicted by the variance in the right set (i.e., $\bar{R} = .59$);
3. Of the redundant variance, 93 percent is associated with the first canonical variate;
4. Despite the large value of $R_{c_2} = .66$, the second canonical variates have very small amounts of variance associated (5 percent in both the left and right sets);
5. The eight canonical factors of the right (and smaller) set extract 100 percent of the variance of that set (which is simply to assert that the smaller set is completely factored in the canonical solution);
6. The redundancy of the right set (student grades) given the left set is $\bar{R} = .61$; and

7. Of the redundant variance of the right set, 92 percent is associated with the first canonical variate.

The utility of $\bar{R}$ is as a summary index. In general it is not to be viewed as an analytic tool. Certain associated indices, however, have obvious analytic applications. For example, the proportion of redundant variance associated with a given factor is instructive in determining whether the factor deserves interpretation and further attention (in the case noted above, a canonical correlation of .66 was associated with only .05 of the variance of either side, and only 4 percent of the redundant variance -- in short, this index instructs us differently than does the canonical correlation alone).

FIGURE I

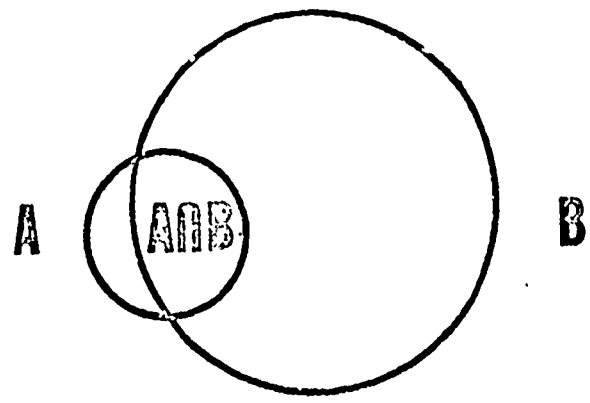

TABLE 1

FACTOR STRUCTURE FOR LEFT SET. COLUMNS ARE CANONICAL FACTORS. ROWS ARE TESTS.

1	-.786	.061	-.082	-.313	.054	.163	-.251	.026
2	-.828	-.163	.018	-.191	-.082	.174	-.276	.031
3	-.707	-.462	.009	-.444	.066	-.102	.018	-.152
4	-.800	-.031	.178	-.095	-.071	.451	-.026	.050
5	-.817	.061	.169	-.194	.003	.311	-.136	-.340
6	-.887	.185	-.096	.074	-.080	.005	-.081	-.005
7	-.917	.119	-.055	-.148	.205	-.016	.120	.050
8	-.836	-.066	.088	-.245	-.046	.082	-.001	.210
9	-.903	-.212	-.086	.099	.083	-.042	.069	-.182
10	-.839	-.351	.016	-.006	-.022	.008	.160	-.136
11	-.752	.048	.581	-.123	.063	.053	-.105	-.113
12	-.798	-.360	.136	.011	.065	-.076	-.243	.096
13	-.726	-.190	.218	-.126	.447	.321	-.198	-.023

FACTOR STRUCTURE FOR RIGHT SET. COLUMNS ARE FACTORS. ROWS ARE TESTS.

1	-.847	-.322	-.065	.094	.212	-.326	-.033	-.119
2	-.795	-.446	-.014	-.067	-.230	.255	.117	-.182
3	-.951	.140	.011	-.108	.095	.046	-.099	.206
4	-.878	.241	-.011	.025	-.194	-.055	-.057	-.354
5	-.901	.127	.315	.227	.080	.073	.093	-.002
6	-.743	.001	.540	-.134	-.189	-.021	-.180	-.263
7	-.800	.027	.088	-.222	.412	-.111	.195	-.288
8	-.727	-.079	.209	.034	.063	.335	-.361	-.416

TABLE 2. Components of Redundancy Measure

L E F T S E T

Factor	I Canonical R R_c	II R-Squared λ	III Variance Extracted VC	IV Redundancy $\lambda \cdot VC$	V Proportion of Total Redundancy
1	.9021	.814	.668	.544	.927
2	.6625	.439	.049	.022	.037
3	.5015	.251	.038	.010	.016
4	.3886	.151	.039	.006	.010
5	.3098	.096	.022	.002	.004
6	.2785	.078	.038	.003	.005
7	.1500	.022	.025	.001	.001
8	.0722	.005	.020	.000	.000

Total Variance Extracted from Left Set = .899

$\overline{R}$, Total Redundancy for Left Set, Given Right Set = .586

R I G H T S E T

Factor	I Canonical R R_c	II R-Squared λ	III Variance Extracted VC	IV Redundancy $\lambda \cdot VC$	V Proportion of Total Redundancy
1	.9021	.814	.695	.566	.923
2	.6625	.439	.050	.022	.036
3	.5015	.251	.056	.014	.023
4	.3886	.151	.018	.003	.004
5	.3098	.096	.045	.004	.007
6	.2785	.078	.038	.003	.005
7	.1500	.022	.030	.001	.001
8	.0722	.005	.068	.000	.001

Total Variance Extracted from Right Set = 1.000

$\overline{R}$, Total Redundancy for Left Set, Given Right Set = .613

REFERENCES

Cooley, William W., and Lohnes, Paul R. Multivariate Procedures for the Behavioral Sciences. New York: John Wiley and Sons, 1962. 211 pp.

Hotelling, Harold. "The Most Predictable Criterion." Journal of Educational Psychology 26:139-142; Feb. 1935.

Lohnes, Paul R., and Marshall, Thomas O. "Redundancy in Student Records." American Educational Research Journal 2:19-23; Jan. 1965.

Meredith, William. "Canonical Correlations with Fallible Data." Psychometrika 29:55-65; March 1964.

Rozeboom, William W. "Linear Correlations between Sets of Variables." Psychometrika 30;57-71; March 1965.

Shannon, Claude E., and Weaver, Warren. The Mathematical Theory of Communication. Urbana: University of Illinois Press, 1949. 117pp.

Attachment C

Announcement of Training Fellowships

AMERICAN INSTITUTES FOR RESEARCH

Institute for Research in Education

Project TALENT Training Fellowships

Beginning September 1, 1966, Project TALENT is offering a postdoctoral program for training in computer and multivariate applications to educational research. Participants will explore a particular area of research using Project TALENT data and participate in the following seminars:

(1) Project TALENT research

(2) Computer applications to educational research

(3) Statistical analysis including multivariate statistics

(4) Research methodology applicable to large-scale educational research

Professional Staff includes:

William W. Cooley, Project Director
Marion F. Shaycoft, Associate Director
Paul R. Lohnes, Director of Guidance Studies
Charles E. Hall, Director of School Studies
Bary G. Wingersky, Director of Computer Systems
Lyle F. Schoenfeldt, Data Bank Coordinator

This program is primarily designed for those who are now holding, or who plan to hold, positions at colleges and universities which involve the training of educational research workers.

Financial support from the Office of Education permits a stipend of $8,500 and relocation costs. Final selection of fellows for academic year 1966-67 will be made on May 31, 1966.

Interested individuals should contact:
William W. Cooley
Director of Project TALENT
135 North Bellefield Avenue
Pittsburgh, Pennsylvania 15213

Project TALENT is a longitudinal study of American high school students which is investigating factors influencing educational and vocational choices. In March 1960 tests were given to 440,000 students in 1,353 secondary schools. These students are being followed up one, five, ten, and twenty years following graduation from high school.