INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.
Preliminary variable selection and data preparation strategies for configural frequency analysis and other categorical multivariate techniques

Tam, Hak Ping, Ph.D.
The Ohio State University, 1992

Copyright ©1992 by Tam, Hak Ping. All rights reserved.
Preliminary Variable Selection and Data Preparation Strategies for Configural Frequency Analysis and Other Categorical Multivariate Techniques

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

by

Hak Ping Tam
The Ohio State University

August, 1992

Dissertation Committee:
Dr. J. J. Kennedy
Dr. T. R. Knapp
Dr. E. A. Stasny

Approved by:
Advisor
College of Education
Department of Education Services and Research
To

My Lord Jesus,
My parents, and
Chao-Hui
ACKNOWLEDGEMENTS

I am indebted to my advisor, Dr. John Kennedy, who has given me complete freedom to pursue my academic interest. His trust, guidance and constant help is highly appreciated. I count myself blessed in having Dr. Thomas Knapp on my committee. His timely arrival as a faculty member in the College of Education has been fortunate for me. I am grateful for his insight, helpful comments, and enthusiasm to read every word on every draft. Sincere thanks also go to Dr. Elizabeth Stasny, who in addition to reading every word of what I have written, is generous with her constructive critique of my ideas. Her comments have contributed significantly to the present work.

I must thanks my parents for their love, care, patience, and expectations. It is for them that I have striven to finish my doctoral studies. To my good friends, Ping-Chak Wong, Shu-Yuan Chang, and Suh-Yuam Hsiao, I note special thanks for the time they offered typing tables, figures and the bibliography. Above all, I would like to thank Chao-Hui Wu for your assistance when I needed it the most. Your friendship have been the most refreshing and cherishable to me.
During my stay at The Ohio State University, I experienced many blessings and abundant graces from my Lord Jesus. I have grown mature in many ways. These have been the best and the most meaningful years in my life so far. I thank you Lord.
VITA

June 29, 1954 ...................... Born - Hong Kong

1973 - 1977 ......................... B.A. in Mathematics, University of California at Berkeley

1977 - 1979 ......................... Graduate Studies in Mathematics, University of California at Berkeley

1980 - 1982 ......................... Mathematics and Physics Teacher, Immanuel Lutheran English Middle School, Hong Kong

1984 - 1986 ......................... Graduate Teaching Assistant, Mathematics Department, The Ohio State University

1985 ......................... M.A. in Education, The Ohio State University

1987 - 1988 ......................... Graduate Research Assistant, College of Education, The Ohio State University


FIELDS OF STUDY

Major Field: Education


Studies in Educational Evaluation. Dr. James Altschuld.
TABLE OF CONTENTS

DEDICATION ......................................................ii
ACKNOWLEDGEMENTS ...........................................iii
VITA ......................................................................v
LIST OF TABLES ..................................................x
LIST OF FIGURES ..................................................xii

CHAPTER PAGE

I. INTRODUCTION ..................................................1
   General Background .........................................5
   Nature of the Problem ........................................6
   Early Development of Categorical Data Analysis
     Techniques .................................................8
   Decomposition of Pearson Chi-square ....................11
   Various Approaches for Multi-way Tables .................13
   Classification Scheme ......................................14
   Considerations to Enhance Henning and
     Rudinger's Taxonomy ....................................15
   Further Considerations ....................................17
   Drawbacks of the Model Fitting Approach ................19
   Data Pre-processing Considerations .......................24
   Configural frequency Analysis as a Data
     Analysis Device .........................................25
   Data Pre-processing in the CFA Context .................26

II. REVIEW OF LITERATURE REGARDING CFA ...............28
   History of Configural Frequency Analysis ...............28
   What Constitutes a Type ? ................................30
   The Formal Setup for CFA ................................30
   Advantages and Disadvantages of CFA .................37
   Identification of Types by Binomial Test ...............39
   Simultaneous Testing in CFA ..............................43
   CFA in Longitudinal Research ............................48
   Modern Development of CFA ..............................49
III. METHODOLOGY .................................................51

Overview .....................................................51
Data Preprocessing in Terms of CFA per SE ..........53
Data Preprocessing by PRU .................................56
Data Preprocessing by Pearson Chi-square Statistic ........................................64
Data Preprocessing by Correspondence Analysis ....66
Data Preprocessing by Stepwise Logistic regression ........................................70
Data Sets Used in the Present Study ....................71
Data Analysis ..................................................77
Comparison .....................................................79

IV. RESULTS .....................................................81

For the First Data Set ......................................81
CFA Approach ..................................................81
PRU Approach .................................................85
Pearson Chi-square Approach ............................88
Correspondence Analysis Approach ....................91
Stepwise Logistic Regression Approach .................94
Summary of Search ...........................................95
For the Second Data Set ....................................96
CFA Approach ..................................................96
PRU Approach .................................................100
Pearson Chi-square Approach ............................103
Correspondence Analysis Approach ....................104
Stepwise Logistic Regression Approach .................106
Summary of Search ...........................................107

V. DISCUSSION ................................................109

General Comments ..........................................109
Discussion on CFA Approach .............................111
Discussion on PRU Approach .............................112
Discussion on the Pearson Chi-square Approach ..........114
Discussion on Correspondence Analysis Approach ........114
Discussion on Stepwise Logistic Regression ...............115
Limitations of Study .........................................116
Recommendations ............................................117
Future Study ..................................................118
APPENDICES

A. Taxonomy of models for qualitative data ........122
B. An introduction to correspondence analysis ....125
C. Correspondence plots for the study of freshmen's alcohol consumption ............132
D. Correspondence plots for the study of productivity of teacher educators .........140

BIBLIOGRAPHY ................................................149
# LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The five variables and their categories chosen for study on freshmen's consumption of alcohol</td>
<td>73</td>
</tr>
<tr>
<td>2. The nine variables and their categories chosen for study on productivity of Ohio's teacher educators</td>
<td>76</td>
</tr>
<tr>
<td>3. Summary table of the five strategies regarding their capability in handling variable selection and category reduction</td>
<td>80</td>
</tr>
<tr>
<td>4. The sum of square of the Z and the corrected Z statistics in CFA for the study of freshmen's consumption of alcohol</td>
<td>82</td>
</tr>
<tr>
<td>5. Configural frequency analysis of the frequency of drink and the greek affiliation variables</td>
<td>85</td>
</tr>
<tr>
<td>6. The PRU analysis of the 4 independent variables for the study of freshmen's alcohol consumption</td>
<td>86</td>
</tr>
<tr>
<td>7. Variables selection by way of Pearson chi-square per degree of freedom for the study of freshmen's alcohol consumption</td>
<td>90</td>
</tr>
<tr>
<td>8. Percentage of total eigenvalue explained by the first two principal axes in the correspondence plot for the study of freshmen's alcohol consumption</td>
<td>93</td>
</tr>
<tr>
<td>9. Order of entry of independent variables as selected by stepwise logistic regression in the study of freshmen's alcohol consumption</td>
<td>94</td>
</tr>
<tr>
<td>10. Summary table of the order of entry with respect to various variable selection strategies</td>
<td>95</td>
</tr>
</tbody>
</table>
11. The sum of square of the Z and the corrected Z statistics in CFA for the study of productivity of teacher educators in Ohio ....97

12. Configural frequency analysis of the number of presentation, bachelor's major in education, and master's major in education variables ........99

13. Using PRU as a criterion in selecting variables and reducing categories for the study of productivity of teacher educators in Ohio ...102

14. Variable selection by way of Pearson chi-square adjusting for the degree of freedom for the study of productivity of teacher educators ..104

15. Percentage of total eigenvalue explained by the first two principal axes in correspondence plot for the study of productivity ..........105

16. Order of entry of independent variables selected by stepwise logistic regression for the study of productivity of teacher educators ..107

17. Summary table of order of entry of variables with respect to the various variable selection strategies .........................107
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Schematic diagram for ( t ) equal from 0 to 3 binary variables</td>
<td>32</td>
</tr>
<tr>
<td>2. Schematic matrix diagram illustrating the mechanism of correspondence analysis</td>
<td>68</td>
</tr>
</tbody>
</table>
CHAPTER I

INTRODUCTION

In any survey conducted in the field of education today, there are, typically, certain items that pertain to the demographics of the respondents. Oftentimes some of the research questions are related to discovering if some of the demographic variables will influence the responses to items of substantive interest in any specific way. Generally speaking, if the variables involved are categorical in nature and the numbers of variables are many, those educational studies of moderate to small sample size will usually result in sparse tables that are problematic to handle. The asymptotic theory behind most of the statistics for such tables may not be valid. Moreover, the more variables included in a model to explain the variation of some dependent variables, the greater the estimated standard error will be. In addition, the model will become more dependent on the observed data (Hosmer and Lemeshow, 1989). With many variables incorporated in the model, the cost of
computation will also be higher. Besides, some of these variables have little effects on the dependent variables. Not only will their presence in the model contribute little towards the understanding of the relationship between the independent and the dependent variables, but may even mask the interpretation of such relationship. Obviously, there needs to be a way to select a meaningful subset of variables and/or to reduce the number of categories for further study.

There are, however, situations where variable selection and category reduction are not desirable. Shapiro (1982), for example, presents an example in which the odds ratios between two dichotomous variables are the same for each level of a third dichotomous variable, but which yield a smaller odds ratio when the levels of the third variable are being collapsed. This phenomenon is known as Simpson's Paradox, which warns of the fact that association between two dichotomous variables may be reversed after collapsing across the third. In this case, the structure of the data as assessed by the odds ratios is changed, which may not be desirable to the researcher. Likewise, in the setting of log-linear analysis, there are times when the marginal association among some variables is larger than its corresponding partial association with respect to a certain set of variables. This external set of variables then constitutes a set of suppressor variables the effect of which may be of interest to the researcher. Under such
circumstances, variable selection procedures will not be appropriate in the sense that the effects of interest may be masked or eliminated in the process of selecting only a limited set of variables.

On the other hand, the usual practice of reducing the number of categories may not be appropriate in certain situations. Baglivo, Olivier and Pagano (1988) present a 2 x 5 table that yields a Pearson chi-square statistic with \( \chi^2(4) = 8.155, p < .05 \). Yet by combining the low count categories, the chi-square statistic changes to \( \chi^2(2) = 0.002, p < 0.99 \). Accordingly, Everitt (1992) advises that combining categories should be avoided for most \( r \times c \) contingency tables. With the advent of fast algorithms and the availability of inexpensive computing power, the remedy is to use the generalized Fisher's exact test that has recently been developed (see Mehta & Patel, 1983, 1986). The algorithm suggested by Baglivo et al. (1988) uses a hybrid of exact and asymptotic techniques and can even extend to multi-way tables. However, their algorithm only pertains to such statistics as the likelihood ratio test, Fisher's exact test, and the Pearson \( X^2 \) test at the present stage of development.

To the extent that the researcher can only handle several variables meaningfully and that he/she is interested in the categories themselves, he/she will need to consider ways that can render the study manageable to the human
minds. This can be done by using data pre-processing that involves two aspects, i.e., by selecting important variables and by reducing categories with respect to certain a priori criteria. It is for this specific consideration that the present study attempts to compare five variable selection and/or category reduction techniques before more elaborated modelling or analytical methods are applied to study two specific data sets.

In recent years, a new technique called configural frequency analysis for categorical data has been devised to identify the existence of groups of respondents who reply with certain distinctive characteristics more or less often than by chance. This technique is even more problematic in relation to the presence of many variables, in the sense that many hypothesis tests for patterns of distinctive characteristics may be unnecessary if some of the variables are not important, according to some a priori criteria, to be included for consideration in the first place.

Thus the intent of the present study is to use configural frequency analysis as a motivation to discuss several data pre-processing procedures that can select subsets of potentially meaningful variables and/or reduce the number of categories per variable. Prior to evaluating different data reduction procedures, a brief discussion of a taxonomy of categorical data analytical techniques is presented.
GENERAL BACKGROUND

One of the most frequently encountered data types in behavioral science research is categorical data. In recent years, much attention has been given to the development of techniques for analyzing this kind of data. McCutcheon (1987, p. 6) explains that one of the reasons for the enthusiasm for renewed development of categorical data methods is the realization that many variables that are useful in theory building — for example the variable indicating traits as used in psychology and syndromes as used in medical fields — are not continuous. Another reason cited by McCutcheon is that over half of the variables available in several of the most widely analyzed data sets in social science are categorical by nature. Similarly, Henning and Rudinger (1985) claim that a growing number of methodologists believe that most variables in psychological theories are categorical "in the sense that ... the majority of hypotheses derived from them express ordinal relations between variables rather than quantitative ones."¹ (p. 297). They maintain that since the prevalence of categorical data is a reality within the realms of social and behavioral science, emphasis should therefore be directed at looking for and developing statistical

¹ What Henning and Rudinger mean by "quantitative" is equivalent to what is commonly referred to as "continuous" by other researchers.
techniques that can handle such data.

NATURE OF THE PROBLEM

In light of the prevalence of, and renewed interest in, categorical data, it is not surprising to witness a remarkable surge of analytical techniques developed for such data in the past two decades. Yet with this sudden proliferation of categorical data techniques, most applied researchers are bewildered by the number of choices made available to them. Not only is it hard to remain current with the state of the art in the development of these analytical techniques, there exist very few guidelines with regard to the choice of one technique over the other according to their relative strengths and weaknesses. It appears that what is much needed is a way of classifying these methods into a framework as a first step to help alleviate this difficulty.

The present study first presents a brief overview of the vast repertoire of techniques that can handle such data. Then this study focuses on a relatively new method known as configural frequency analysis, or CFA in short, that originated in West Germany in the late 1960s and the early 1970s and is designed to identify whether there are distinct groups of individuals with unique patterns of characteristics and/or whether their specific response patterns are showing up purely by chance. Useful as this
technique may be, it becomes impractical when the number of possible response patterns gets beyond the manageability of the researcher. Hence the main intention of the present study is in exploring and empirically comparing ways to perform data preprocessing in terms of selecting only important variables that are responsible for the highest amount of variation in some other categorical variable so as to render the performance of CFA more effective.

This thesis is organized into five chapters. The first chapter is devoted to an explanation of the nature of the problem together with a broad perspective on the statistical methods designed especially for categorical data. The methods will be presented within a taxonomy so as to illuminate the relationships among these methods. Afterwards, a formal introduction to configural frequency analysis will be presented and a particular disadvantage of this method, namely the futility in management in case of too many configurations, will be discussed. Chapter two is devoted to a literature review of the development of configural frequency analysis. Chapter three delineates several potential data preprocessing techniques in terms of variable selection and category reduction so as to facilitate a more effective configural frequency analysis. These techniques will then be applied to two different data sets for empirical comparison. The results of this comparison will be reported in chapter four. The concern is
to perform an exploratory study into the relative merits of the different preprocessing ideas in terms of which variables are selected and which categories are combined, as well as the interpretability of the results. These comparisons will be discussed in chapter five, where suggestions for further direction of study will also be proposed.

EARLY DEVELOPMENT OF CATEGORICAL DATA ANALYSIS TECHNIQUES

Generally speaking, there are three types of variables that can give rise to categorical data: data derived from discrete variables; data from continuous variables that are discretized into categories; and, finally, data from latent continuous variables that may be manifested in terms of two or more categories (e.g. attitude).

At the beginning of the twentieth century, the only statistical technique available for the analysis of categorical data was the chi-square test as developed by Karl Pearson in 1900. The Pearson chi-square test was originally developed for the test of independence of variables in a two-way contingency table. A significantly large value of the chi-square statistic would suggest the rejection of the null hypothesis that there is no association between the two variables in the population from which the sample(s) was(were) drawn. With or without the
rejection of the null hypothesis, a wide variety of measures of association were gradually introduced, dating back to the beginning of the present century, for the purpose of measuring the strength of the relationship between the two variables (Goodman and Kruskal, 1954). Such measures can be regarded as descriptive statistics even if no test of independence is performed (Agresti, 1990).

A major disadvantage of the Pearson chi-square statistic is that it is dependent on the sample size from which it is calculated. Thus, by itself, it cannot be used to measure the strength of the relationship. Moreover it does not indicate the direction of a relationship. To circumvent the former drawback, a common practice is to divide the Pearson's chi-square statistic by the sample size (Hildebrand, Laing & Rosenthal, 1977). The square root of the new statistic thus created is called the phi coefficient or the root mean square contingency coefficient, which is not limited just to 2 x 2 tables (see e.g. Upton, 1978; Wickens, 1989). Yet it still suffers the same limitation as the chi-square statistic in that it provides no information about the direction of the relationship. Hildebrand et al. regard the latter statistic as a measure of distance, indicating "how far" the probability structure is from statistical independence (p. 46).

On the other hand, there exist some research problems that, by their very nature, require investigation
This hypothesis investigates whether or not there exist equal distributions of proportions among the categories of one variable across the categories of the other variable. This hypothesis requires a product multinomial sampling scheme where individuals are sampled from fixed marginal totals. It is different from testing independence in a contingency table where the total sample size is fixed, thereby constituting a multinomial sampling scheme. It turns out that the same Pearson chi-square statistic can also be used to test this hypothesis. The interpretation, of course, will be different since a different sampling scheme is used and a different question is being asked in this case.

Many of the problems encountered in behavioral science research, however, involve more than two variables, thus rendering the Pearson chi-square test inappropriate as a tool for analyzing multi-way tables. Even though the Pearson chi-square statistic can be extended to higher dimensional tables to test the null hypothesis of total independence of the variables involved, other tests must be used to assess additional hypotheses of interest. Besides, a significantly large value of the chi-square statistic only implies that the null hypothesis of total independence has to be rejected. As for the source of departure from the null, the test itself says nothing about it (Lienert &
Wolfrum, 1980). Many earlier applied researchers, for example in sociology, tried to get around this problem by breaking up the multi-way tables into a series of two-way tables and then applying the Pearson chi-square test to each table. Such procedures are prone to the escalation of cumulative type I errors, to say nothing about the cumbersomeness of calculations before the computer age, as well as the difficulty in extrapolating the interpretation of two-way tables to the intricate network of relationships among variables in a multi-way table.

DECOMPOSITION OF PEARSON CHI-SQUARE

Another way to handle the complexity of multi-way tables is to adopt the additive decomposition method of the overall chi-square statistic as suggested by Irwin and further developed by Lancaster (Irwin, 1949; Lancaster, 1949, 1950). Given a set of p variables, the basic idea behind decomposition is to express the Pearson chi-square for the association among the p variables as a function of the sum of the pth order interaction term together with the set of (p-1)th order association terms, adjusting for the fact that there is an overlapping component among the (p-1) variables. Each of the (p-1)th order association terms can likewise be decomposed into sums of the interaction terms with lower order association terms. The typical procedure is to calculate the corresponding chi-square statistic for
each order of the association term and then either compare the resulting statistics to some critical values with respect to the appropriate degrees of freedom at a pre-specified significance level or compute the p-values for decision making.

There are, however, major disadvantages to the decomposition method proposed by Lancaster and other researchers that limit its usefulness as applied to multi-way data. First, the sequential chi-square testing procedures mentioned above are most likely to be dependent on each other. Thus the actual significance level is usually higher than the nominal \( \alpha \) significance level. In order to minimize the commission of a type I error, a common practice is to employ the well-known Bonferroni adjustment procedure. Rather than using \( \alpha \) as the significance level, \( \alpha' = \alpha / t \) is used, where \( t \) is the total number of significance tests made. Another problem that is related to the first one concerns the independence of observations on the \( p \) variables. Lancaster's method of chi-square decomposition is based on the assumption that there is total independence among the observations. However this condition is seldom checked by practitioners who are responsible for demonstrating it substantively. Moreover, it is possible that variables that are deemed independent substantively may turn out to be statistically correlated with each other. On the contrary, variables that are statistically uncorrelated
may be dependent by virtue of the nature of the problem (e.g., in a repeated measures design).

VARIOUS APPROACHES FOR MULTI-WAY TABLES

Since the late 1950's, interest in categorical data has flourished. As a result, much new ground has been explored with regard to the analysis of multi-way contingency tables. Such well-known techniques as log-linear analysis (Bishop, Fienberg and Holland, 1975; Goodman, 1978; Grizzle, Starmer and Koch, 1969; Haberman, 1974) and logistic regression (Cox, 1970), together with such less well-known but increasingly popular techniques as correspondence analysis (Benzécri, 1973; Greenacre, 1984), configural frequency analysis (Krauth and Lienert, 1973), information theory (Kullback, 1959), Galois lattices (Andor, Joó, & Mérö, 1985), partial order scalogram analysis (Shye, 1978), qualitative canonical analysis (Kendall and Stuart, 1967), factor analysis of dichotomous variables (Muthén and Christofferson, 1981; Muthén, 1992), latent class variable analysis (Lazarsfeld, 1950), discrete discriminant analysis (Gilbert, 1968), polychoric correlation models (Olsson, 1979), and cluster analysis for categorical data (van Buuren and Heiser, 1989) have been launched, thus constituting a battery of new techniques for the study of multi-way frequency data.
CLASSIFICATION SCHEME

There are perhaps many plausible ways to sort related methods into groups. The attempt, in general, is to assemble together those methods that are based on the same statistical approach. Such an effort for organizing categorical data into a taxonomy was discussed in Bentler (1980). The context in which he worked was within causal modelling in the study of cognitive development, an area in which categorical data are relatively abundant. More recently, Henning and Rudinger (1985) further expanded Bentler’s taxonomy to include additional statistical techniques, yet still within the framework of modelling which they perceived to be appropriate for hypothesis-testing of cognitive and developmental data. Whereas Bentler provided a concise survey of the purposes and suggestions for application of the various methods, Henning and Rudinger presented an adapted and slightly expanded table of the taxonomy without much further exposition. Their table has been reproduced with permission from the publisher as Appendix A for ready reference.
One of the weaknesses with Appendix A is that the classification basically concerns techniques that are mostly catered for modelling and is not the only approach to organize a taxonomy. Only a few of the listed methods, for example correspondence analysis and dual scaling, pertain to data analysis rather than model fitting or formulated to study causal relationships. Moreover, Appendix A is not exhaustive in the sense that there exist some relatively new techniques that do not fall into the above classification scheme. Hence, further efforts to render the various methods into a more inclusive taxonomy are warranted.

CONSIDERATIONS TO ENHANCE HENNING AND RUDINGER'S TAXONOMY

First, there are those techniques that are based on some kind of regression for the variables involved. These methods make use of variations of the least-squares approach to handle categorical data. As contrasted with the maximum-likelihood method, which is oriented towards parameter estimation, the least squares method is more oriented towards prediction (Wold, 1983). It includes such techniques as the Grizzle, Starmer and Koch (GSK) method that builds on weighted least squares, the partial least squares (PLS) method and the analysis of table (ANOTA) method that utilizes ordinary least squares (Keller et al., 1985), and the polychoric correlation method (Olsson, 1979) that is based on generalized least squares. On the other
hand, the log-linear approach as well as the logistic regression approach are based on maximum likelihood estimation.

Another type of analysis may concern itself with the scaling of a set of objects according to the attributes they possess as indicated by which categories they belong to among the variables making up the contingency table. Correspondence analysis is an example of this type of technique. Canonical analysis for categorical data is another example, since it has been shown formally to be equivalent to correspondence analysis (de Leeuw, 1973). Moreover, it can also be demonstrated that principal components analysis as modified to apply to qualitative data is also equivalent to correspondence analysis (Pielou, 1984). Other examples include factor analysis for dichotomous data and latent trait models (Rasch, 1960).

Still another type of analysis may use the idea of residual analysis for the identification of possible sources of deviation of the observed data in a contingency table from the hypothesized model. Such well known methods as log-linear analysis and logit analysis fall into this category. However, there are techniques that follow more or less the same philosophy but concern themselves specifically with deviations from the independence model. Examples include the chi-square analysis, quasi-independence method, and the Mantel-Haenzel method (1959). On the other hand, methods
like the chi-square automatic interaction detection, CHAID (Kass, 1984) and nested logit analysis (Magidson, Swan, and Berk, 1981) "in which a contrast or group of contrasts is supposed to be operative at some levels of a variable (but not at all levels)" (see Clark, Deurloo, & Dieleman, 1988, p. 202).

Also there is a class of methods that capitalize on the algebraic structure and logical order relations of the data. They focus on the interconnections between the variables themselves rather than on the overall fit of some models or on the generation of certain indices. The resulting knowledge structures are then graphically presented. Galois lattices and partial order scalogram analysis belong to this setting.

FURTHER CONSIDERATIONS

The considerations above is far from being unique, nor will it yield an exhaustive and complete classification scheme. Each of the techniques mentioned above is highly developed. Some may even contain several statistical features and thus are capable of being cross-classified. It is hence very natural that there exist several other alternative ways to re-classify these techniques. For example, a new scheme is to classify according to the purposes of the researcher behind the usage of the above methods, much in line with but expanded from those suggested
by von Eye (1990a, p. 4). If the purpose of the researcher is in modelling the observed frequencies as being accounted for by some latent factors, then such methods as factor analysis for dichotomous variables, latent class variable modelling and polychoric correlation modelling will form a group by themselves. If the purpose is to explore the structure of relationships among the variables that constitute the multi-way table, then such methods as log-linear analysis, logistic regression, canonical correlation for categorical data, and information theory are typical techniques to use. If the purpose is to study patterns of characteristics of the groups of subjects as demonstrated by the attributes they possess among the categories of the variables, then the researcher should consider using methods like configural frequency analysis and prediction analysis of cross-classifications. Finally, if the researcher desires to investigate the relationship among the categories of the variables by clustering the categories into some kind of groupings, then the researcher might consider techniques like correspondence analysis, dual scaling, or other kinds of scaling methods.

Owing to the existence of so many categorical data techniques, there is certainly a need to further compare and contrast them in order to establish guidelines regarding the circumstances under which it is more appropriate to use a certain method.
DRAWBACKS OF THE MODEL FITTING APPROACH

Undeniably, the past two decades have witnessed a soaring trend of searching for and fitting models, e.g., log-linear models, to data collected within the realm of social and behavioral science research. Apparently, many applied researchers think that good research consists of searching for a good model that can provide an acceptable fit to the observed data and justifying it by supplying a reasonable explanation of the structure, especially a causal one, at hand. Yet it must be cautioned that neither model fitting nor causal modelling should not be looked upon as the panacea for analyzing data. Causal modelling has, nevertheless, its own weaknesses that are frequently overlooked by most researchers. The Journal of Educational Statistics has devoted a whole issue to the pros and cons of causal modelling, although the discussion centered around interval data rather than categorical data.

First, all models carry with them certain assumptions about the observed variables and possibly the latent variables as well. At times such assumptions can be very strong. For instance, the sample size involved in the study may be too small and thus hamper the pertinence of the first-order asymptotic statistical results that are frequently relied upon in modelling procedures. Another example is the identification problem that must first be examined before any parameter estimation can be considered.
These technicalities imply that whenever a researcher is testing the fit of a model, concern must be raised regarding the robustness of the model against plausible violations of the underlying assumptions. This concern also constitutes the first and foremost criticism by Freedman (1987) on casual modelling. Thus practitioners should verify and report the validity of the underlying assumptions before formally taking the modelling approach. In most circumstances, it is advisable to perform more exploratory data analyses whenever possible prior to any rigorous attempt at causal modelling.

Secondly, there is the problem of misspecification in the form of omitting variables that are relevant to the model that characterizes the data. Specification errors then refer to the existence of discrepancies between the model attempted and the "true" model. Rather than merely expanding the model by including more variables without justification, the only salvation to the misspecification problem is to identify variables that are based on either sound theory or fact.

Thirdly, it has been common experience that oftentimes several models may fit the data equally well. The present causal modelling methodology for selecting the "right" model depends largely on some kind of chi-square goodness-of-fit test, even though there are advocates for using various descriptive fit indices together with residual
analysis and cross-validation techniques to help in such
decision making (Bonett and Bentler, 1983; Pedhazur, 1982).
The choice of the right model can be subjective at best when
several competitive yet different models yield compatible
goodness-of-fit statistics. In this case, any model adopted
may suggest a different understanding of the problem. In
this sense, causal modelling is more subjective and
exploratory than what some proponents expect it to be. As a
matter of fact, the choice of a model is always subjective,
at least in the setting of social sciences.

Moreover, the chi-square test is sample dependent in
the sense that it runs into problems when the sample size is
either too small or too large. In the case of a small
sample size, the resulting contingency table will contain a
lot of empty cells or cells with low frequencies. The chi-
square statistic thus calculated will then be biased, and
also violate the $\chi^2$ distribution assumption which is an
asymptotic result requiring a large sample size. Hence this
statistic may not be very useful. Furthermore, there is the
complication of distinguishing structural zeros (where it is
theoretically impossible to have any observations in the
cells) from sampling zeros (where cells are empty by chance
owing to sampling errors) and the subsequent statistical
adjustment will differ in the two cases. For instance, the
frequency of Tibetan Christians will probably be zero in a
survey. However, whether this cell should be treated as a
structural zero or sampling zero will require some judgment call. On the other hand, the larger the sample size, the more likely the chance of achieving a significant test statistic. As a result, more restricted models are easily rejected whenever the sample size is large. Thus it can be quite difficult to obtain a parsimonious model that fits the data well.

One further problem with the casual modelling approach is related to the narrowness of the recent trend in causal modelling in which researchers rely exclusively on the type of information thus generated to answer their research questions. Oftentimes, they are told by others, say journal editors, that they need to perform log-linear analyses for their data. However, such results can at times be quite limited in scope or even misleading. A better way to choose the type of analysis to perform is to focus on the problem and decide on the best method of analysis and carry it out accordingly. Subsequently, depending on the research problem, it is advisable to obtain a judicious combination of different kinds of information derived from different analytical approaches, thereby leading to insights in the structure of the data that cannot be gained otherwise (Kim, 1984). The taking of the-best-of-two-worlds approach has been advocated, for example, by van der Heijden and de Leeuw (1985) to use correspondence analysis to find meaningful structure in data, and also log-linear analysis to estimate
the parameters as well as to understand the relationships among the variables. The last, but certainly not the least, disadvantage of the causal modelling approach is that sometimes the research problem may call for treatment of data other than model fitting. To apply causal modelling irrespective of the underlying research problem is deemed most inappropriate in any research endeavor. At this point it should be emphasized that there is a difference between the data analysis of contingency tables and the sequential choice of models that fit the data. It is not uncommon for researchers to pay more attention to the models they are searching for than to the underlying contingency table. And while in the process of model fitting, most practitioners mechanically apply the tour-de-force of the mechanism of their modelling technique, for example in model searching within log-linear analysis, and do not pay much attention to the variables and the kind of data they are presently dealing with. As a result they suffer from two major drawbacks in their studies as pointed out by Clark, Deurloo and Dieleman (1988). First, they may have too many independent variables, each of which may have several categories, that have significant relationships with the dependent variable. The cross-tabulation of all these variables will thereby constitute a sparse table that results from "thinning" out the sample data amongst the
cells. The parameter estimates thus obtained are at best dubious in nature if not totally unreliable. Besides, such huge tables are not manageable as far as interpretation of the effects is concerned.

To remedy this situation, most practitioners will subjectively choose several variables they are particularly interested in and may even collapse some of the categories in order to make it amenable to the modelling technique. Such attempts, however, will usually cause the studies to suffer from the second drawback, namely, including relatively unimportant variables in their models.

DATA PRE-PROCESSING CONSIDERATIONS

From the above discussion, it seems clear that there is a need for developing some kind of objective data pre-processing procedure in order for a meaningful and reliable data analysis procedure to be found. A useful data pre-processing technique will select the most important independent variables, according to a certain criterion, from a larger set and subsequently reduce the number of categories once a variable is selected without losing much information from the data. In the past, several researchers, including Higgins and Koch (1977), Green (1978), and Magidson (1982) have discussed the benefit of pre-processing prior to the use of logit analysis on the data. According to their experiences, Clark et al. (1988)
recommend, from among several possible techniques, two good choices, namely, the proportional reduction in uncertainty, or PRU, together with the chi-square automatic interaction detection, or CHAID. These are applicable in situations, e.g. when the sample size is small, that require the selection of a few independent categorical variables prior to the study of their effects on the dependent categorical variables and at the same time relinquish those other independent variables that have relatively weaker effects.

CONFIGURAL FREQUENCY ANALYSIS AS A DATA ANALYSIS DEVICE

In some psychological or educational studies, for example in the scrutiny of personality traits, the goal of the researcher is to examine the distinctiveness of people with different combinations of personality types rather than establishing any causal structure. It is in this setting that a relatively new technique, namely, configural frequency analysis, or CFA, has been developed. CFA may be used to investigate patterns of characteristics among various groups of subjects. It has wide application potentiality and has been used successfully in defining clinical syndromes in the psychopathological fields (e.g., Krauth and Lienert, 1975) in a cross-sectional perspective as well as in the study of trends from a longitudinal viewpoint (e.g., Lienert and Oeveste, 1988).

CFA is a non-parametric multivariate statistical
technique that is intended for identifying types and syndromes within a group of subjects and for analyzing the interaction structure among the attributes observed in the subjects. More specifically, this method concerns itself with identifying groups of individuals who manifest similar response profiles on two or more categorical variables significantly (in the statistical sense) more often than expected under the null hypothesis of independence among the categorical variables. It can be applied to underlying models other than the independence model as well.

However, when there are many variables involved, the interaction structure can be very complicated. The number of hypotheses being tested will be large, thereby opening up the risk of committing cumulative Type I errors.

DATA PRE-PROCESSING IN THE CFA CONTEXT

Because of this disadvantage, it is also desirable to apply data pre-processing procedures within a CFA context so as to select only important variables to be included in the study as well as to reduce the number of relevant categories within each variable selected. Such modifications will render the results in CFA more interpretable and also reducing the pressure of requiring an immense sample size, thereby improving the applicability of CFA. In addition, with a smaller number of variables studied, there will be fewer cells possible and, hence, the
research problem is more manageable. But before any
discussion of data preparation methods, a rather
comprehensive review of the literature regarding CFA is
necessary so as to set a context in using CFA as a
motivation for considering several variable selection and/or
category reducing techniques. This is the matter of concern
in the next chapter.
CHAPTER II

REVIEW OF LITERATURE REGARDING CFA

HISTORY OF CONFIGURAL FREQUENCY ANALYSIS

Configural frequency analysis, or CFA, was originally proposed by Lienert in 1968 in a lecture, and later published in 1969, as a tool for detecting the existence of types and syndromes among a group of subjects from a set of multivariate categorical data collected about them. Today, it has been applied in a wide variety of fields, including psychology, psychiatry, medicine, sociology and education. The early papers on this relatively new technique were written exclusively in German, thus rendering this technique uncommon among applied researchers in the English-speaking world. One of the first presentations of CFA in English was published in 1975 (Lienert and Krauth, 1975). More recently, there appeared many papers on the modifications and applications of CFA that were written in English. However, most of these studies appeared in the Biometrical Journal, which is published in Germany and has drawn the attention of only a
specific audience among the English-speaking practitioners. Hence up to the present moment, it is very unfortunate that most of the researchers writing on this technique are confined to those of European origin. Yet with the recent publications of textbooks that introduce CFA in great detail (e.g. von Eye, 1990a; Kennedy, 1992), there is hope that this technique can get more attention among non-European practitioners.

The earliest form of CFA was essentially an eye-balling technique. It was later developed into an inferential technique by Krauth (Krauth and Lienert, 1973). Prior to the advance of configural frequency analysis as a statistical technique, the most common practice was to use factor analysis to look for types of traits among a group of subjects. A variant form known as Q-technique can be used to cluster the subjects rather than the variables. The major concern, however, against the utilization of factor analysis is that it is based on correlation matrices and makes use of the first-order interactions of the variables (Krauth, 1985). As a result, it cannot capture those types that are caused solely by higher-order interactions. On top of this are the methodological problems related to the mechanism of factor analysis, such as the determination of the number of factors and the indeterminacy problem, about which different researchers have different approaches. Besides, there are several measures of similarity and
dissimilarity that can compete with the Pearson correlation coefficients used in factor analysis. Tversky (1977) pointed out that the types identified by these procedures may not correspond to the types that the practitioner has in mind from an intuitive point of view.

WHAT CONSTITUTES A TYPE?

According to Krauth (Krauth, 1985, pp. 162-3), the basic concept of type as used in CFA is based on the following intuitive assumptions:

1. Types are identified by attributes that are observed among the subjects.
2. A type is identified whenever a researcher repeatedly finds subjects in his sample who show the same pattern or configuration of attributes.
3. Any attribute that is continuous or has many values (i.e. a variable that is either interval, or ordinal, or nominal with many categories by nature) is discretized into a categorical variable with few categories.
4. A certain dependence of the attributes is necessary for an observed configuration to be accepted as a type. This would exclude those spurious configurations of categories that are formed by lumping certain attributes which together will already form a type with some unnecessary categories from other variables. The type thus formed from the larger assembly of categories is admissible only by the virtue of the presence of the essential attributes.

THE FORMAL SETUP FOR CFA

Using the multinomial sampling scheme, consider a sample of individuals who are characterized by whether they
possess the various attributes corresponding to a set of $t$ categorical variables, i.e. $\{V_i: 1 \leq i \leq t\}$. Furthermore, suppose each variable $V_i$ consists of $s_i$ mutually exclusive categories that are to be denoted by $C_{i,1}$, $C_{i,2}$, $\ldots$, $C_{i,s_i}$ etc. Here the first subscript, namely, $i$, refers to the $i$-th variable, while the second subscript refers to the order of the categories for the $i$-th variable. In general, an arbitrary category will be represented by $C_{i,j_i}$ with $1 \leq j_i \leq s_i$ for $i$ ranging from 1 to $t$. Thus there can be a total of $s_1 \times s_2 \times \ldots \times s_t$ possible combinations of categories across all variables. Each possible combination, in CFA terminology, is regarded as a configuration, and can be generally represented as a $t$-tuple, $(C_{1,j_1}, \ldots, C_{t,j_t})$. The observed frequency in this configuration is denoted by $f_{j_1 \ldots j_t}$. The counts of occurrences across all the possible configurations follow a multinomial distribution. A schematic diagram for the case of $0 \leq t \leq 3$ dichotomous variables is presented in Figure 1 on the next page. The symbol $N$ in the figure refers to the total sample size.
| t (number of | t-dim.   | # of marginals | # of | proportions to sample size |
| binary variables) | table |  | configurations | frequencies |  |
|------------------|-------|----------------|------|---------------------------|
| 0                | 0-way (a cell) | □ | 1 | 2^0=1 | N | 1 |
| 1                | 1-way (a tuple) | ☐ ☐ ☐ | 2+1=3 | 2^1=2 | f=(f_1,f_2) | p=(f_1/N,f_2/N) |
| 2                | 2-way (4-fold) | ☐ ☐ ☐ | 2+2=4 | 2^2=4 | f=(f_{11},f_{12},f_{21},f_{22}) | p=(f_{11}/N,f_{12}/N,f_{21}/N,f_{22}/N) |
| 3                | 3-way table | ☐ ☐ ☐ ☐ | 4+4+4=12 | 2^3=8 | f=(f_{111},f_{112},f_{121},f_{122},f_{211},f_{212},f_{221},f_{222}) | p=(f_{111}/N,f_{112}/N,f_{121}/N,f_{122}/N,f_{211}/N,f_{212}/N,f_{221}/N,f_{222}/N) |

Figure 1. Schematic diagram for t equal from 0 to 3 binary variables
Consider a typical $r \times c$ contingency table. For a test of independence of the row and column variables, the null hypothesis will be:

$$H_0 : p_{ij} = p_{i+} p_{.j} ,$$

where $p_{ij}$ is the probability of an observation in the $(i,j)$ configuration, and $p_{i+}$ and $p_{.j}$ are the $i$th row and $j$th column marginal respectively. The Pearson chi-square test for independence can be regarded as the weighted sum of the squares of distances between the observed frequencies and the corresponding expected frequencies under the hypothesis of independence. In this case, the weight is equal to the reciprocal of the expected frequency in the respective cell. Its formula is represented as below:

$$X^2 = \sum_{i,j} \frac{(f_{ij} - e_{ij})^2}{e_{ij}} \quad \ldots \quad (1)$$

Thus if the overall Pearson chi-square statistic turns out to be significantly larger than expected under the model, those components that have very large contributions to $X^2$ would suggest that their corresponding cells may be quite different from the other cells, indicating the sources of departure from the null.

In the early stage of development of CFA (Lienert, 1969; Krauth and Lienert, 1982), while adopting the common practice of residual analysis, the individual component of the Pearson chi-square has been regarded as a squared normal
variate by the originators of CFA, or equivalently, as being distributed as chi-square with one degree of freedom. However, they have presented no justification for this treatment other than claiming that it is a heuristic way. The present author attempts to supply, in the following, a plausible rationale for this "heuristic" approach and then discusses the appropriateness of this method as a way to identify cells with frequencies that differ from what is expected under the null.

For motivation, instead of the multinomial sampling scheme assumed earlier, it is assumed that a Poisson sampling scheme has been followed. Thus each cell is an independently generated Poisson variate with both mean and variance equal to the Poisson parameter $\mu$ (Wickens, 1989, p. 135). Hence the standardized variable $z$ is given by the following formula:

$$z = \frac{f_{ij} - E(X)}{\sigma_x} = \frac{f_{ij} - \mu}{\sqrt{\mu}} \quad \ldots (2)$$

and is traditionally regarded as a standardized residual. If the standardized variable turns out to be significantly larger than that expected from the underlying null model, then over- or under-representation has occurred for that particular configuration. If the null model happens to be the independence model, then there is evidence that the variables involved are not independent as presumed under the
null.

However, each $z$ as defined above is only approximately normally distributed with a mean of zero and unit variance (Upton, 1978; Andersen, 1990). Andersen also points out that the residuals calculated this way do not even have the same variance. Furthermore, the residuals among the various cells will always be related to each other, thereby violating the independence assumption in the Poisson sampling situation (Upton, 1978). Rittich (1988) claims that the approach discussed above is very conservative, meaning that it is very weak in terms of identifying over- or under-representation of cell frequencies.

In order to endow this cell-wise identification of over- or under-representation with an inferential feature, Krauth and Lienert (1973) proposed to assume a binomial distribution in each cell test instead of the chi-square components to identify the sources of departure from the null. More specifically, they assume each observation of a configuration constitutes an independent trial of a Bernoulli experiment that is an independent realization of a configuration. Thus there are $N$ (where $N$ is the total sample size) independent trials in total, with each observation resulting either in a particular configuration or some other (von Eye, 1990a).

The probability $p_{ij}$ of any given configuration is
given by $e_{ij}/N$ under the null. The probability of having $f_{ij}$ observations in the $(i,j)$ configuration can be found by assuming a binomial distribution for that configuration. In other words, the probability of observing a frequency of $f_{ij}$ is given by:

$$P(f_{ij}) = \left( \frac{N}{f_{ij}} \right)^{f_{ij}} q_{ij}^{(n-f_{ij})}, \quad \ldots \quad (3)$$

where $q_{ij} = 1 - p_{ij}$. In order to test the null hypothesis, the cumulative probability of the $(i,j)$ configuration with frequencies equal to $f_{ij}$ or higher is calculated and its value is compared to the nominal $\alpha$. In symbols, the cumulative binomial p-value is given by:

$$F(f_{ij}) = \sum_{k=f_{ij}}^{N} \left( \frac{N}{k} \right) p_{ij}^{k} q_{ij}^{(n-k)} , \quad \ldots \quad (4)$$

If $F(f_{ij})$ is less than $\alpha$, or equivalently $f_{ij} > Np_{ij}$, the $(i,j)$ configuration, in CFA terminology, constitutes a type. If, however, the observed frequency turns out to be significantly less than the expected frequency, then an antitype is said to have been identified. In this case, the lower and upper limit of the summation in equation 4 are modified to the following terms:

$$F(f_{ij}) = \sum_{k=0}^{f_{ij}} \left( \frac{N}{k} \right) p_{ij}^{k} q_{ij}^{(n-k)} , \quad \ldots \quad (5)$$
Thus in general, those subjects who belong to the cells with large deviations can be regarded as appearing together more or less often than may be expected by chance.

In classical CFA, one typically proceeds from comparing the observed data to the expected frequencies under a base model, usually the total independence model, in order to identify the structure of relationships among the variables by means of locating over- or under-represented cells with respect to the base model. In this setting, a type is a configuration that occurs with a higher frequency as compared to the expected values under the base model. On the contrary, an antitype is a configuration that occurs more rarely than is expected. Accordingly, the aim of CFA is two-fold, namely, to determine if types and/or antitypes exist at all in the multi-way tables and, in case they do exist, to identify the corresponding configurations as such (Victor, 1989).

ADVANTAGES AND DISADVANTAGES OF CFA

Krauth (1985) summarizes three of the frequently suggested reasons to promote CFA. Firstly, the definitions of types and antitypes are intuitively appealing to researchers especially in psychology, education and related fields. Secondly, CFA is computationally less intensive and less dependent on complicated computer packages than other
multivariate techniques. More importantly is the feasibility in CFA to allow researchers to interpret the configuration directly without having to check various assumptions, as in most parametric techniques, or having to deal with latent structures whose existence can be very dubious.

Furthermore, von Eye (1990a) suggests that CFA can be used supplementary to log-linear modeling, especially when the latter method fails to verify a theoretical model that the researcher expects will fit the data. Thus even though the association structure cannot be substantiated, the patterns of characteristics of the subjects can still be identified and studied. An even more informative approach is that adopted by Kennedy (1992) in which CFA is utilized to study the residuals after fitting a predetermined log-linear model to the data. This treatment has the further advantage of enabling the researcher in understanding how the fitted model differs from the data.

One final advantage of CFA, as pointed out by Krauth (1982) is its flexibility in being easily modified to study different problems. A detailed delineation of such versatility in terms of applications can be found in von Eye (1990a).

One of the principal disadvantages of CFA is its impracticality when the problem involves numerous
categorical variables, each carrying with it numerous categories. In this situation, the number of configurations possible will become unmanageable. In other words, the interpretation of the configurations will become quite difficult. In addition, the sample size required under this setting will be very large. One other disadvantage of CFA is the large number of tests that need to be done when the number of configurations is large. This may result in the escalation of type I error rates. However, this has been traditionally controlled by using such corrective procedures as Bonferroni's or Holm's procedure for this purpose.

The last, but certainly not the least, disadvantage is the intrinsic weakness of CFA's framework in that over-representation in some configurations will likely imply under-representation in some other configurations (Victor, 1989; Victor and Kieser, 1991). In other words, the cells identified as types and antitypes in the binomial setup are dependent upon each other. This latter disadvantage will be discussed further under the topic of modern development of CFA in the present chapter.

IDENTIFICATION OF TYPES BY BINOMIAL TEST

One early line of research in CFA involved the searching for a better statistical test for determining if a configuration constitutes a type or an antitype congruent
with the above assumptions. In Lienert's original conceptualization (1969), the first test employed was the Pearson chi-square test. However this practice has a serious drawback, as already discussed above. Subsequent development was to use the binomial test as an exact test for comparing the observed to expected frequencies under the null (Krauth and Lienert, 1973). Nevertheless, as pointed out by von Eye (1990a, pp. 18-19), there are two reasons why the binomial density function should be approximated by other methods to avoid technical computational difficulty, especially more so in the days when computer technology is not so developed. First, for larger sample sizes, the calculation of the factorial terms in the binomial equation will be very tedious. Even more important is that the factorial terms will be immense and cause arithmetic overflow under certain language compilers for some micro- or even mini-computers. This constraint has, instinctively, led to much research that dealt with searching for a better approximation method.

In their first introduction to the English-speaking world, Lienert and Krauth (1975) mentioned four possible approximation methods. The first one was the well known standard normal approximation to the binomial distribution by means of the following formula:
This rendering is appropriate if \( N \), the sample size, is large and \( p_{ij} \), the probability of the obtained frequency for each configuration is not too small. The version they used in their paper was the one with a continuity correction, which adjusts for the fact that a continuous distribution is used to approximate the discrete binomial distribution. The second method they mentioned was to use the chi-square approximation if the expected frequencies of most, say 80\%, of the configurations exceeded 5. This approximation is relatively easy to compute. Moreover, there is research that indicates that the bigger-than-5 rule can be relaxed (see Wise, 1963; Larntz, 1978).

The other two rules they included are seldom used in the CFA literature. These included the Poisson approximation, which is appropriate when \( N \) is large and \( p \) is very small, and the relatively unknown Camp-Paulson approximation, which is basically an F-distribution (see Molenaar, 1970). The approximation is done via the following formula:

\[
F = \frac{(f_{ij} + 1)(1 - p_{ij})}{(N - f_{ij})p_{ij}}, \quad \ldots \ (7)
\]

where \( F \) is evaluated with \( 2(N - f_{ij}) \) degrees of freedom for
the numerator and 2(f_1^j+1) degree of freedom for the denominator.

In order to avoid conservatism in the chi-square approximation above, Lehmacher (1981), and Lehmacher and Lienert (1982) resorted to using asymptotic tests that were derived from the hypergeometric distribution of observed cell frequencies. The basic assumption for this approach is that the margins of the contingency table are fixed before the data are collected. Lehmacher provided evidence that this approach is more powerful than the binomial test. Küchenhoff (1986) further modified Lehmacher's approach by incorporating Yates' continuity correction for the observed frequencies. Küchenhoff substantiated his approach by providing empirical evidence that demonstrated the reduction of non-conservative decisions in comparison to the corresponding approach without the continuity correction. Further evidence regarding the usefulness of the corrected Lehmacher procedure was cited in Rittich (1988).

Heilmann and Schütt (1985) acknowledged that, in terms of computation, Snedecor's F-distribution is the most convenient way to approximate the discrete binomial distribution, since it has a functional relationship to the latter. In their study, they provided elaborate tables for binomial testing via the F-distribution. However, they only reported those values that are in the extreme positive end
of the F-distribution.

In addition, von Eye and Bergman (1986) discussed using Sterling's formula to approximate the binomial distribution. Their study, together with the one performed by von Eye (1990a), adopted an empirical approach by comparing the different tests proposed above for their performance on an artificial set of data. Based on the dual criteria of higher power and fewer non-conservative decisions, it appears that, generally speaking, Küchenhoff's modification on the Lehmacher test performs the best.

SIMULTANEOUS TESTING IN CFA

In the search for types and antitypes within the framework of CFA, numerous hypothesis testings of types and antitypes are performed simultaneously, thereby necessitating extra caution in interpreting the results. von Eye (1990a) discussed two major issues related to simultaneous testing performed on the same set of data in a CFA setting. The first one concerns the mutual independence of multiple tests. What this means is that after a test for type/antitype is found to be statistically significant, the subsequent tests are very likely to be found significant owing to the underestimation of the actual p-values. CFA is particularly inflicted with this problem, since all configurations are tested in a typical CFA endeavor. The
second but associated problem is related to multiple testings, which, because of the numerous tests being effected, some of them would probably turn out to be significant purely by chance.

To remedy for these drawbacks, one possibility is to adopt Goodman's (1965) suggestion which amounts to, in the general setting, the use of the degrees of freedom of the whole contingency table for each chi-square test as a means for alpha control. One of the earliest remedies in the context of CFA was suggested by Krauth and Lienert (1973), and cited in von Eye (1990b), to use Bonferroni's procedure to control for the inflation of the experimentwise type I error rate. However, Bonferroni's procedure, even though more effective than Goodman's suggestion, still tends to be very conservative.

To further improve on a search for types and antitypes, Lehmacher and Lienert (1982) suggested the adoption of the sequential Bonferroni procedure as devised by Holm (1979) for alpha adjustment. The advantage of Holm's approach is that instead of fixing a constant alpha level, as in the case of Bonferroni's approach, for sequential comparison in search of types and antitypes, it takes into consideration the number of tests already performed. In CFA, Holm's procedure is essentially composed of two steps. The first step is to arrange all the p-values
of the approximate binomial tests for all the configurations in an ascending order. The second step is to compare each successive p-value with the adjusted alpha level which is determined by the following formula:

\[ \alpha^* = \frac{\alpha}{(t - m + 1)}, \quad \ldots \quad (8) \]

where \( t \) is the total number of tests (or the number of configurations made) and \( m \) takes on values from 1 to \( t \). Here the second step is essentially reducing the denominator of \( \alpha^* \), i.e. the criterion, by one each time a hypothesis is being rejected, "so that tests can be conducted at successively higher significance levels" (Shaffer, 1986, p. 826).

The decision rule is then to compare each successive p-value with the corresponding adjusted alpha level. If the former is less than the latter, then a type or an antitype has been identified and the search continues with the next p-value. The procedure terminates upon encountering a p-value larger than its corresponding adjusted alpha. Obviously, Holm's procedure is equivalent to Bonferroni's procedure for the first test but it becomes less conservative starting with the second test. The disadvantage of this method, as pointed out by Perli, Hommel and Lehmann (1985), is that it is insensitive to the special structure or relationship that may exist in the set
of hypotheses under test. As a result, Holm's procedure can still be quite conservative.

In view of the possible existence of mutual dependence amongst multiple tests, Hommel, Lehmacher, and Perli (1985) proposed a modified version of Holm's procedure that "is based on the fact that hypotheses on single cells can, under certain conditions, be depicted as intersections of m other cell hypotheses" (von Eye, 1990a, p. 39). Thus for a two-dimensional table with t cells (or t hypothesis tests), it can be shown that if m cell hypotheses, under the condition \( t > m \geq t-3 \), cannot be rejected, then the remaining \( t-m \) hypotheses also cannot be rejected. Moreover, if \( m = t-5 \) cell hypotheses cannot be rejected, then there must be at least one more hypothesis that cannot be rejected. Relying on these properties, Hommel et al. (1985) derived the following set of alpha levels in which several cells can be tested at the same alpha level for types/antitypes. Specifically, the first hypothesis is tested against \( \alpha/t \), the second to the fifth hypotheses against \( \alpha/(t-4) \), the sixth and the seventh one against \( \alpha/(t-6) \), the eighth one against \( \alpha/(t-7) \), etc., until the last hypothesis is tested against the nominal alpha itself.

In addition, according to von Eye (1990a), several recently developed multiple comparison techniques can also be applied in a CFA context. These techniques include the
modifications of Holm's procedure as suggested by Shaffer (1986), which essentially pertains to the logical estimation of the maximum number of hypotheses that can be true, under the condition that m-1 hypotheses have already been rejected. Since all the pairwise hypotheses concern comparisons between means, this maximum number also corresponds to the number of pairs of means that can be equal, given the condition that m-1 previous pairs have already been determined to be different. These numbers can then be used to adjust the alpha levels for the sequence of hypotheses tested. However, von Eye is not specific as to how the Holm-Shaffer procedure, which deals with means, could be applied to tests for types and antitypes in CFA. Moreover, the mechanism of this modification of Holm's procedure does not lend itself to the calculation of confidence intervals.

Furthermore, von Eye suggested that the strategy proposed by Holland and Copenhaver (1987, 1988) can also be applicable if certain dependence assumptions are met. First, Holland and Copenhaver defined the following function, C(x), which is defined for a positive, real-valued variable x:

\[ C(x) = 1 - (1 - \alpha)^{\frac{1}{x}}. \tag{9} \]

This function can easily be shown to have the property that
C(x) ≥ α/x for x ≥ 1. By setting x = t - m + 1, with t and m specified as in the case of Holm's procedure, C(t-m+1) will then be greater than α/(t-m+1). Holland and Copenhaver then replaced the term on the right hand side of Holm's procedure with the larger term, C(t-m+1). In this way, more hypotheses may be rejected in comparison with Holm's procedure, thereby deeming the Holland and Copenhaver procedure more powerful than the latter procedure. Nevertheless, Holland and Copenhaver admitted that the potential gain in power may be quite minimal, since the function C(x) is only slightly bigger than the term α/x.

CFA IN LONGITUDINAL RESEARCH

Technically speaking, the analysis of longitudinal data is relatively more difficult than the analysis of cross-sectional data because of the existence of repeated measurement of the same subjects over time, thereby leading to an artificial increase in sample size through multiple countings (von Eye, 1990b). Equally important is the problem that the samples are no longer independent of each other, since they are comprised of the same group of subjects measured at different points of time. When all the variables involved are categorical in nature, the traditional way of handling longitudinal data is to cross-tabulate the variable measured at time one with the variable
measured at time two etc. into a contingency table. Yet in this setting, classical CFA is not adequate in identifying types or antitypes directly because of the problem of dependent samples. However, if the research interest is in the trend of changes in the variables, then through suitable transformation of the combination of categories across time, various approaches of applying CFA to this type of data are still possible.

Under this circumstance, CFA is concerned with identifying cases with patterns that are either stable or show specific changes over time. von Eye (1990a) summarized its application in the study of increases and/or decreases of the response variable over time, shifts in location over time, time series of different lengths, and various other designs. Since this topic is not directly related to the present study, interested readers are referred to von Eye (1990a, 1990b) for details.

MODERN DEVELOPMENT OF CFA

An initial criticism of CFA is from Victor (1981) who charged that singular types or antitypes may not exist under the null hypothesis of independence since the over-representation of one cell will imply an under-representation somewhere in a multi-way table (see Lienert, 1989). Furthermore, Victor and Kieser (1991) mentioned that
since CFA is basically a residual analysis technique that typically uses the independence model as a basis, then large residuals may indicate the inadequacy of the base model but may not guarantee the existence of types/antitypes because of "masking" effects (Victor and Kieser, 1991). What this means is that for a configuration that has been identified as a type, some of the variables that constitute it may not be the source of influence that contribute to the occurrence of the type. A smaller set of variables may already suffice. The duality concept as well as the consideration of the adequacy of the base model eventually led Victor to pioneer his quasi-independence CFA, which he explicitly formulated in 1988 and which was subsequently expounded on by Lienert (1989). Rather than a cell-wise search, his approach specifically incorporates the possibility of the existence of types in the estimation process (Victor, 1989). The quasi-independence expected frequencies are calculated through iterative procedures as delineated in Victor (1983). Kieser and Victor (1991) further developed a two-stage procedure that can test, at the same time, the existence of types together with the adequacy of the base model.
CHAPTER III

METHODOLOGY

OVERVIEW

Obviously, there are many hypothesis tests for types for a typical CFA problem that involves several variables with each being composed of several categories. In order to control for unnecessary tests that may return statistically significant results purely by chance, caution must be employed to safeguard researchers from reaching conclusions that are without ground. One of the best ways is to test only those hypotheses that the researcher is really interested in based on the literature or for theoretical reasons. Another good way is to conduct data pre-processing to select only important variables, based on pre-determined statistical criteria, to be included in CFA. There are several reasonable schemes that can contribute towards this end. Although exploratory data analysis (EDA) is an important approach to data analysis in its own right, apparently EDA procedures that pertain to categorical data
have no direct bearing on the selection of variables and in the reduction of categories that are of interest in the present study.

The main portion of the present study focuses on several feasible data preprocessing techniques that can be applied to the data prior to the carrying out of a CFA. A total of five procedures are identified in this chapter. They are, namely, data preprocessing by:

1. CFA per se
2. proportional reduction of uncertainties
3. the Pearson chi-square statistic
4. correspondence analysis, and
5. stepwise logistic regression.

The second method, the proportional reduction of uncertainties, which is based on information theory and is not very well-known and hence will be presented in some detail. The rest of the methods will, however, only be briefly explained. A succinct mathematical derivation of correspondence analysis has also been included in Appendix B for ready reference. Subsequently, each of these techniques is applied to two real data sets so as to assess their relative merits in selecting variables and in the reduction of categories.
While considering the identification of a type, Krauth (1982) mentions the possibility that not all variables that constitute a type are responsible for the generation of the type itself. There may be single variables that do not interact with the other variables, and should thus be identified and eliminated. He further suggests a so-called hierarchical CFA procedure through which a feasible way to systematically search for types can be done by first applying CFA to the combinations of categories with all of the variables included. The second step is to search for types for each of the possible combinations with all but one of the variables. This searching procedure is continued by dropping variables one by one until a CFA is performed for each combination that entails only two variables. The final decision is to admit that particular configuration that yields the most clear-cut type/antitype. However, he has not pointed out the criteria with regards to which the most clear-cut type can be defined. Furthermore, this particular approach requires numerous hypothesis tests for types, the total number of which, e.g. for a study with t variables, can be calculated by the following formula:
\[
\binom{t}{1} + \binom{t}{t-1} + \binom{t}{t-2} + \ldots + \binom{t}{2} \\
= 1 + t + \frac{t(t-1)}{2} + \ldots + \frac{t(t-1)}{2} 
\] \quad \ldots \quad (10)

A slight variation to the above procedure but more specifically laid out is the one mentioned in Lienert and Krauth (1975), where a CFA is first performed with all the variables included. Afterwards, all normal approximations to the exact binomials of each configuration are then squared and added together. This sum is then evaluated as a chi-square with its number of degrees of freedom assessed through the following rule:

\[
\text{d.f.} = \text{the number of configurations} - \text{the number of variables involved} - 1 \quad \ldots \quad (11)
\]

The reason why the number of degrees of freedom is computed this way instead of using the regular \((\text{row} - 1) \times (\text{column} - 1) \times \ldots \times (\text{layer} - 1)\) formula is because the above rule will yield larger degrees of freedom and thus can better account for the number of configurations when it is large. With a larger number of configurations, the sum of squares of the \(z\) statistics may be large not because of the presence of large deviations in some configurations but just because more terms are being added together. Hence with larger degrees of freedom, the tests for types and antitypes can be more rigorous thereby minimizing the risk of capitalizing on chance.
Next, CFA's are performed on combinations of categories from all but one variable and the corresponding sums of squares of the unit normal deviates are assessed with the corresponding numbers of degrees of freedom determined using the same rule as above. This procedure of taking away one variable at a time is continued until only bivariate combinations are left for examination by CFA using the same approach as above. Finally, out of all the sums of squares of unit normal deviates computed, the CFA which results in the highest sum of squares is selected, thereby eliminating the variables that are unnecessary in defining configural types.

Some of the disadvantages of the CFA approach are quite obvious. While Krauth's strategy is straightforward, it does not rely on any other information that may render the selection of variables more effective. It should also be pointed out that the order of the procedures suggested by Lienert and Krauth is not important since the final aim is to compare the relative sizes of the sum of the $Z^2$ statistics from the various CFA studies. Furthermore, Kennedy (1992) pointed out that the approximate $Z$ statistic only accounts for the probability of the upper tail of the binomial distribution. Hence he suggested a corrected $Z$ statistic that accounts for the probability at the lower tail in the presence of an antitype. In addition, both of
the above approaches are quite impractical when the number of variables involved in the study is large. Lastly, and probably the most significant weakness of the CFA approach, is that the $Z$ statistics of the various configurations within each CFA study are related to each other. Hence the sum of $Z^2$ does not follow the chi-square distribution as assumed in Lienert and Krauth. As a result, it can, at best, be used only as an index and no inferences should be attempted about it.

Thus the main purpose of the present study is to explore alternative schemes that may seem to be quite promising in contributing towards the goal of categorical variable selection. Their relative merits in performance will be assessed through empirical comparison on two real data sets and then compared to that of the CFA approach. These alternatives are briefly discussed in the following sections.

DATA PREPROCESSING BY INFORMATION THEORY

Researchers have long been interested in studying the strength of the relationship between two variables in a contingency table. Numerous coefficients or indices have been formulated to serve as measures of association. Because of the deficiencies in the chi-square statistic, measures that are based on the chi-square statistic will not
be very promising.

For tables that are made up of nominal variables, a useful measure is the proportional reduction in error, or PRE, which was devised by Goodman and Kruskal (1954) and is not connected to the chi-square statistic. It assesses association by means of comparing the probabilities of making misclassification errors under two prediction rules. For instance, in predicting to which row category a subject belongs in a contingency table, two rules can be used. The first one does not make use of any information from the column categories. This is, in statistical terms, equivalent to independence between the row and column variables. The second rule assumes that the two variables are related and takes advantage of this information in predicting the subject's row classification. Then the reduction in error in using rule two as opposed to rule one can be calculated as follows:

\[
PRE = \frac{P(1) - P(2)}{P(1)}, \quad \ldots \quad (12)
\]

where \( P(1) \) refers to the probability of misclassification error under rule one and \( P(2) \) as the probability of misclassification error under rule two (see Reynolds, 1977). The possible values of PRE range between 0 and 1, with a value of 0 indicates that the probability of misclassification is the same under both rules while a value
of 1 indicates that knowledge of the column categories results in the exact prediction of row categories.

One special case of the PRE is the proportional reduction of uncertainty, or PRU, where the error probabilities are defined in terms of information theory. According to Shannon (1948), the average uncertainty or entropy, \( U \), in a variable, say \( x \), is defined as follows:

\[
U_x = \sum_i p_i \ln \left( \frac{1}{p_i} \right) = -\sum_i p_i \ln p_i. \tag{13}
\]

Here, \( p_i \) refers to the proportion of the observations that belong to the \( i \)th category, and \( \ln \) refers to the natural logarithm function\(^2\). By definition, \( p_i \ln(p_i) \) is set equal to zero for \( p_i \) equal to zero. \( U_x \) is a function of both the number of categories and the distribution pattern of frequencies among the categories. For a given number of categories in a variable, its value is maximized when the distribution is uniform among all categories.

Traditionally, the uncertainty measure can be regarded as a measure of variance of a categorical variable (e.g. Schafer, 1980; Haberman, 1982; Kim, 1984).

\(^2\) Kim (1984) recommended the use of natural logarithms instead of the usual logarithm to base 2, as originally employed by Shannon (1948) and later adopted in communication research, by suggesting its relative convenience to use as well as its use in the likelihood ratio chi-square test.
In a more general context, for a contingency table involving variables \( x \) and \( y \), and following the notation of Brown (1975), let \( p_{ij} \) denote the proportion of observations in the \((i,j)\) cell such that \( \sum_j p_{ij} = 1 \). Also let \( p_{i+} = \sum_j p_{ij} \), and \( p_{+j} = \sum_i p_{ij} \). In this context, the average uncertainty for variable \( x \) can be expressed as follows:

\[
U_x = \sum_i p_{i+} \ln \left( \frac{1}{p_{i+}} \right) . \quad \ldots (14)
\]

On the other hand, the overall expected uncertainty for the \((i,j)\) cell is defined as:

\[
U_{xy} = \sum_i \sum_j p_{ij} \ln \left( \frac{1}{p_{ij}} \right)
= -\sum_i \sum_j p_{ij} \ln p_{ij} . \quad \ldots (15)
\]

According to this definition, it is clear that \( U_{xy} \) is the same as \( U_{yx} \). Using the fact that \( U_x = \sum_i p_{i+} \ln(1/p_{i+}) \) and \( U_y = \sum_j p_{+j} \ln(1/p_{+j}) \), the conditional uncertainty of \( x \) given \( y \) and that of \( y \) given \( x \) are designated as follows:

\[
U_{x|y} = U_{xy} - U_y
= \sum_i \sum_j p_{ij} \ln \left( \frac{1}{p_{ij}} \right) - \sum_j p_{+j} \ln \left( \frac{1}{p_{+j}} \right)
= \sum_i \sum_j p_{ij} \ln \left( \frac{p_{ij}}{p_{i+}} \right) , \quad \ldots (16)
\]

and
The measure \( U_{y|x} \) can be regarded as the uncertainty in classifying a subject in category \( j \) of variable \( y \) given the category in which variable \( x \) falls. It will be at its minimum (i.e. equal to zero) if and only if \( y \) is a deterministic function of \( x \).

One further concept that is useful in the present study is the so-called transmitted information between variable \( x \) and \( y \). It is denoted by \( M(x,y) \) and is specified as below:

\[
M(x,y) = U_y - U_{y|x} = \sum_j p_{j|x} \ln \left( \frac{1}{p_{j|x}} \right) - \sum_j \sum_{j'} p_{jj'} \ln \left( \frac{p_{j|x}}{p_{j'|x}} \right). \tag{18}
\]

It can be shown that \( U_y - U_{y|x} \) is the same as \( U_x - U_{x|y} \). Thus \( M(x,y) \) can be regarded as the mutual information between variable \( x \) and \( y \). If \( y \) is dependent on \( x \), \( M(x,y) \) can be thought of as the variation in \( y \) that is explained by \( x \).

Using these measures of uncertainty, Kim (1984) defines the PRU index for a two-way table as:

\[
U_{y|x} = U_{xy} - U_x = \sum_j \sum_{j'} p_{jj'} \ln \left( \frac{1}{p_{jj'}} \right) - \sum_j p_{j|x} \ln \left( \frac{1}{p_{j|x}} \right) = \sum_j \sum_{j'} p_{jj'} \ln \left( \frac{p_{j|x}}{p_{j'|x}} \right). \tag{17}
\]
For ease of computation, the PRU index can be re-expressed as follows:

\[ r_{yx}^2 = \frac{(U_y - U_{y|x})}{U_y} \]  \hspace{1cm} \cdots \text{(19)}

This index serves as a measure of association for categorical variables. It indicates the "relevancy rather than significance of the relationship" between the two variables (Clark et al., 1988, p. 204). It reflects the gain in explanation of one variable with knowledge about another variable. The index will be at its minimum value, i.e. zero, if and only if \( y \) is independent of \( x \), and at its maximum value, i.e. one, if and only if \( y \) is a function of \( x \). Note that this index is asymmetrical in the sense that \( r_{yx}^2 \) is different from \( r_{xy}^2 \), since they carry different terms in their denominators. Kim (1984) further suggests a symmetrical version of PRU which is incorporated into such major statistical software packages as SPSS, SAS, and BMDP. He has, moreover, extended PRU to a multivariate setting.

It should also be noted that PRU is related to the likelihood ratio chi-square statistic, which is generally
denoted by $G^2$ or $L^2$ in the literature. It can be shown that the transmitted information between $x$ and $y$, $M(x,y)$, is related to the likelihood ratio chi-square statistic, $L^2$ according to the following formula:

$$M(x,y) = \frac{L^2}{2N}, \quad \ldots \quad (21)$$

where $N$ refers to the total sample size. Moreover, Kim (1984) shows that $r_{yx}^2$ can be expressed as a ratio of two likelihood ratio chi-square statistics.

There are some researchers, e.g. Costner (1965), and Clark, Deurloo, and Dieleman (1986), who strongly prefer the use of proportional reduction of error (PRE) criterion, the interpretation of which does not depend on the distribution of the underlying data, to measure the strength of association between variables. Note that proportional reduction in uncertainty (PRU) is one such measure, it is related to the likelihood ratio chi-square, and its magnitude is independent of the sample size. Therefore its use as a measures of association has much to be recommended.

Clark et al. (1986, 1988) applied the PRU index for the purpose of variable selection. For a given dependent variable, they use PRU in a sequential manner to first select the independent variable that yields the highest PRU with the dependent variable. The categories of this selected variable are combined in a pairwise manner and the
PRU measure of the two-way tables thus formed with the dependent variable are assessed one by one. The combination of categories with the lowest decrease in PRU is then selected. In general, the lower the decrease in PRU, the lower the reduction in the explained variation there is in the dependent variable.

As a stopping rule, Clark et al. (1988) refer to the likelihood ratio chi-square statistic, $L^2$, as a guideline. However, they emphasize that this statistic is used as a subsidiary to the PRU index. While the categories are combined in a pairwise manner, the decrease in $L^2$ between the original table and each of the reduced tables is assessed and compared to the critical $\chi^2$ distribution values according to the difference in the number of degrees of freedom for the two tables. (Clark et al. used a 1% significance level in their work.) If this test results in significance and if, at the same time, there is a substantial decrease in the PRU, then the particular combination of categories in consideration should probably not be attempted. Furthermore, the $L^2$ statistic can also be used as a supplement to the PRU in selecting variables. If the increase in $L^2$ between the original table and the table with an extra variable turns out to be significant, then the new variable might be considered as a potential explanatory variable.
Taking this new combined variable as a starting point, further pairwise combinations of categories are attempted using the smallest change in uncertainty in the dependent variable as the criterion. The combining process continues until there is a drastic drop in the PRU values. Clark et al. (1986) call the independent variable thus selected the reduced category variable. The remainder of the independent variables are selected and combined in like manner.

Clark et al. (1986) claim that this procedure can result in reasonable subsets of cross-classifications and that it will "produce a substantial amount of explanatory value while avoiding thin data in the cells" (p. 769).

DATA PREPROCESSING BY THE PEARSON CHI-SQUARE STATISTIC

Another potential data preprocessing technique devised by Clarke and Koch (1976), and again employed by Higgins and Koch (1977), makes use of the Pearson chi-square statistics adjusted for their numbers of degrees of freedom as a means to select variables. The justification behind this approach is the fact that the expected value of a chi-square statistic between two categorical variables is equal to its number of degrees of freedom under the null hypothesis of no association. Thus the stronger the relationship between the variables, the larger the chi-
Clarke and Koch proceed by first cross-tabulating the dependent variable with each of the independent variables in a series of two-way tables. Pearson chi-square statistics are calculated for each of the tables and then divided by their corresponding numbers of degrees of freedom. The first explanatory variable chosen is the one with the biggest chi-square per degree of freedom. This chosen variable is then combined in sequence with each of the remaining independent variables. The combined cross-classifications formed in this manner are then joined with the dependent variable to form generalized two-way tables. Pearson chi-square statistics adjusted for their degrees of freedom are again calculated and the same decision rule as described above applies. This procedure continues subject to the following stopping rule.

The selection procedure terminates in accordance with the results of two types of statistics. The first statistic amounts to adding up the Pearson chi-square statistics that are calculated for the two-way tables formed by the variable in consideration for inclusion with the dependent variable controlled for each level of combinations of the variables that have already been selected. Their corresponding numbers of degrees of freedom are, likewise, summed up. If the resulting statistic is significant, then
the variable in consideration will be selected. However, if the statistic is not significant, the selection procedure will be terminated.

Clarke and Koch (1976) note that the above stopping rule has a disadvantage in that it loses its usefulness owing to the thinning of data after the first few steps of the selection process. Hence they suggest the use of the Mantel-Haenszel chi-square, or, whenever appropriate, the multiple degrees of freedom extension of it, as an alternative statistic. It possesses a chi-square distribution with the appropriate number of degrees of freedom and it represents the effect of the independent variable under consideration on the dependent variable, over and above all the combinations of the variables already selected. Since the alternative statistic makes use of information across all combinations of the variables already selected, it does not suffer from the thinning of data. Also, if this statistic fails to reach statistical significance at, say, the .05 level, the variable selection procedure will be terminated.

DATA PREPROCESSING BY CORRESPONDENCE ANALYSIS

Correspondence analysis is a multivariate descriptive analysis technique that recently has generated interest increasingly as a tool for data exploration. It
was developed independently by different researchers in various disciplines, each using a slightly different perspective. The geometric perspective, as embodied in the French school under Benzécri, regards it as a tool to find the best simultaneous representation of several categorical variables, each with several categories, on a low-dimensional graph (Lebert et al., 1984). This feature of visualizing the data carries with it an appeal for empirical researchers as they analyze data.

One possible approach for formulating correspondence analysis is to scale simultaneously the data matrix by means of the reciprocal of the square root transformation of both the row and column marginals. The resulting matrix is then multiplied by its own transpose matrix, and the new product matrix is then subjected to eigenvalue decomposition. Finally the coordinates of both the row and column categories that are associated with the low dimensional graph can be found by transforming the eigenvector matrix that results from the above-mentioned eigenvalue decomposition using the reciprocal of the square root of the row and column marginals, respectively. A more mathematical presentation of the mechanism of correspondence analysis from another perspective has been included in Appendix B. A schematic diagram that illustrates the preceding procedure is presented on the next page.
Figure 2. Schematic matrix diagram illustrating the mechanism of correspondence analysis
Moreover, each of the axes of the corresponding plot is associated with a percentage that represents the proportion of variance explained by the sum of all the eigenvalues by that axis. According to Jambu (1991), interpretation of the plot should be preceded by giving "a specific meaning to each axis and to the proximities between points when projected onto the factor axes." (p. 189) Generally speaking, it is dangerous to interpret the relative proximity of two points with one representing a row category and the other a column category in the graph. It is, however, appropriate to interpret the distances among the points that represent the row categories alone or the column categories alone (see e.g. Greenacre, 1984; Lebart, Morineau, and Warwick, 1984). Under these circumstances, if the two rows or columns have similar profiles, they can be regarded as being similar to each other. Meanwhile, profiles of the row and column marginals are usually projected onto the origin of the correspondence plot. As a result, when a row or column category is in the proximity of the origin, the profile of this category is similar to that of the row or column marginals.

Hence it is proposed in the present study to take advantage of the visual features of correspondence analysis to combine those categories that are relatively close to one another as a way to achieve category reduction. This can
also be done in a stepwise manner by first performing correspondence analyses on two-way tables formed by combining each of the independent variables with the dependent variable. The results are then represented on two-dimensional correspondence plots. The independent variable that yields a pair of axes that together explain the highest percentage of total eigenvalue of the respective two-way contingency tables with the dependent variable will be selected first. Meanwhile, its categories are examined for closeness and subjective judgment is used to determine whether certain categories can be combined. If the answer to the latter question is positive, then the categories closest together are combined to achieve category reduction. This selected and reduced variable is then combined with each of the remaining independent variables. The resulting joint cross-classifications are then joined with the dependent variable to form generalized two-way tables and the same variable selection and category reduction procedure is repeated.

DATA PREPROCESSING BY STEPWISE LOGISTIC REGRESSION

Logistic regression is appropriate when assessing the effects of several independent variables on a dependent variable that is categorical in nature (Agresti, 1990). A stepwise procedure analogous to the one used in multiple
linear regression is also available for logistic regression (see e.g. Christensen, 1990). Even though stepwise logistic regression cannot be used to reduce categories, it is included in this study as a viable way of selecting categorical variables. Its results can be compared with those methods suggested above and their relative merits discussed.

DATA SETS USED IN THE PRESENT STUDY

The first data set used in this study is based on a survey of 434 freshmen in a small liberal arts college in the mid-west. The mean age of freshmen in the sample is 17.9 years old with a range from 16 to 20 years of age. The instrument used is the Core Alcohol and Drug Survey form. One of the variables of interest in this study is the number of times a subject consumed more than four drinks at a sitting. This variable is regarded as an ordinal variable with five levels. The research interest is to find out if gender, grade point average, age at which the respondent first used alcohol and whether or not the respondent lives in a fraternity or sorority will in any way be related to his or her amount of alcohol consumption. Here the gender variable is a nominal variable with two levels, the Greek affiliation variable consists of two levels, grade point average is an ordinal variable with thirteen levels, and age
at first use of alcohol is also an ordinal variable with nine levels. Technically speaking, since the intervals are not the same in between adjacent categories, even though the latter two variables have numerous categories, they are still regarded as ordinal rather than interval variables. As a whole, there are as many as 2,160 cells in the five-way contingency table for the data set. The frequency distribution of the data set is presented in the following table on the next page. The numbers of missing values are 16 for the dependent variable, 15 for gender, 29 for Greek affiliation, 23 for grade point average, and 20 for the age at first consumption variables, respectively.
Table 1. The five variables and their categories chosen for study on freshmen's consumption of alcohol.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Categories</th>
<th>Frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent variable</td>
<td>0 none</td>
<td>197</td>
</tr>
<tr>
<td></td>
<td>1 one</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>2 two</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>3 three</td>
<td>66</td>
</tr>
<tr>
<td>Frequency of drinks at a sitting</td>
<td>4 ≥ four</td>
<td>41</td>
</tr>
<tr>
<td>Independent variables</td>
<td>0 male</td>
<td>173</td>
</tr>
<tr>
<td></td>
<td>1 female</td>
<td>246</td>
</tr>
<tr>
<td>Gender</td>
<td>0 no</td>
<td>112</td>
</tr>
<tr>
<td></td>
<td>1 yes</td>
<td>293</td>
</tr>
<tr>
<td>Greek affiliation</td>
<td>0 F</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1 D -</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2 D</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>3 D +</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>4 C -</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>5 C</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>6 C +</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>7 B -</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>8 B</td>
<td>120</td>
</tr>
<tr>
<td></td>
<td>9 B +</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>10 A -</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>11 A</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>12 A +</td>
<td>28</td>
</tr>
<tr>
<td>Variables</td>
<td>Categories</td>
<td>Frequencies</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Age at first use of alcohol</td>
<td>1 never</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>2 &lt; 10</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>3 10-11</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>4 12-13</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>5 14-15</td>
<td>162</td>
</tr>
<tr>
<td></td>
<td>6 16-17</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>7 18-20</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>8 21-25</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>9 26 +</td>
<td>1</td>
</tr>
</tbody>
</table>

In the context of CFA, the research question is to investigate the combinations of characteristics that constitute either types or antitypes with regard to alcohol consumption. There are, however, so many configurations that both variable selection and category reduction are warranted for a more effective CFA.

The second data set originated in a study of the characteristics of prospective teacher educators and their reasons for wanting to become teacher educators. The focus of interest in the present study concerns, however, the level of scholarly and research productivity of graduate students who are specifically enrolled in the program of teacher education. The target population consists of all students who enrolled in advanced graduate programs that purported to prepare teacher educators in the state of Ohio. There are eight universities in Ohio that offer Ed.
D. or Ph.D. programs in teacher education. All eight of them set a high priority on research in their programs. Altogether there were about seven hundred prospective teacher educators during the 1991-92 academic year, of whom 568 provided responses that can be included in the present study. A complete description of these incipient teachers of teachers is described in the doctoral research of McCullough (1992).

The dependent variable used in the present study assesses the number of national, regional, and state professional presentations given by the respondents while in the doctoral program. It is an ordinal variable with five levels, ranging from none to four or more presentations. Eight independent variables are included in the study. These have been identified, on an a priori basis, to be potential predictors of the scholarly productivity of the respondents. They include such variables as: whether the students have completed a master's thesis (nominal, 2 levels), their grade point averages (ordinal, 3 levels), whether they majored in education at the bachelor's level (nominal, 2 levels), whether they majored in education at the master's level (nominal, 2 levels), their age (ordinal, 3 levels), whether they are enrolled as full-time students (nominal, 2 levels), the number of years they taught full-time in K-12 (ordinal, 6 levels), and lastly, when they decided to
become teacher educators (ordinal, 7 levels). The frequency distributions for the variables in this data set are presented in the table below. The numbers of missing responses are 2 for the dependent variable, 2 for the Master education major, 5 for enrollment, and 13 for the time of decision to become teacher educator variables, respectively. Altogether there are 40,320 cells in this data set.

Table 2. The nine variables and their categories chosen for study on productivity of Ohio's teacher educators.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Categories</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent variable</td>
<td>0 none</td>
<td>269</td>
</tr>
<tr>
<td></td>
<td>1 one</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>2 two</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>3 three</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>4 four or more</td>
<td>120</td>
</tr>
<tr>
<td>Presentation</td>
<td>Master thesis written</td>
<td>1 yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 no</td>
</tr>
<tr>
<td></td>
<td>GPA</td>
<td>1 ≤ 3.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 &gt; 3.0 to ≤ 3.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 &gt; 3.6</td>
</tr>
<tr>
<td></td>
<td>Bachelor education major</td>
<td>1 yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 no</td>
</tr>
</tbody>
</table>
### Variables

<table>
<thead>
<tr>
<th>Categories</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Master education major</td>
<td></td>
</tr>
<tr>
<td>1 yes</td>
<td>433</td>
</tr>
<tr>
<td>2 no</td>
<td>133</td>
</tr>
<tr>
<td>Age at Ph.D.</td>
<td></td>
</tr>
<tr>
<td>1 ≤ 31</td>
<td>139</td>
</tr>
<tr>
<td>2 32 to 37</td>
<td>158</td>
</tr>
<tr>
<td>3 38 to 42</td>
<td>129</td>
</tr>
<tr>
<td>4 &gt; 42</td>
<td>142</td>
</tr>
<tr>
<td>Enrolled</td>
<td></td>
</tr>
<tr>
<td>1 full time</td>
<td>233</td>
</tr>
<tr>
<td>2 not full time</td>
<td>330</td>
</tr>
<tr>
<td>Teaching experience</td>
<td></td>
</tr>
<tr>
<td>0 none</td>
<td>119</td>
</tr>
<tr>
<td>1 &lt; 2 years</td>
<td>24</td>
</tr>
<tr>
<td>2 2-5 years</td>
<td>105</td>
</tr>
<tr>
<td>3 6-9 years</td>
<td>96</td>
</tr>
<tr>
<td>4 10-13 years</td>
<td>60</td>
</tr>
<tr>
<td>5 &gt; 13 years</td>
<td>164</td>
</tr>
<tr>
<td>Decision to become teacher educator</td>
<td></td>
</tr>
<tr>
<td>1 before high school</td>
<td>33</td>
</tr>
<tr>
<td>2 during high school</td>
<td>32</td>
</tr>
<tr>
<td>3 undergraduate</td>
<td>72</td>
</tr>
<tr>
<td>4 after bachelor's degree</td>
<td>68</td>
</tr>
<tr>
<td>5 graduate study</td>
<td>112</td>
</tr>
<tr>
<td>6 after master's degree</td>
<td>178</td>
</tr>
<tr>
<td>7 doctoral degree</td>
<td>60</td>
</tr>
</tbody>
</table>

**DATA ANALYSIS**

Both of the data sets included for analysis embody the asymmetric mode of inquiry framework, meaning that they both concern the effects of several independent variables on a single dependent variable.
The first data set represents a simpler study in that it involves only five variables, of which four are independent and the remaining one is the dependent variable. However, two of the variables have many categories, some of them being of zero frequency count. Hence when these variables are crossed to form contingency tables, some of the categories turn out to be so problematic as to create cells with zero expected values with reference to their corresponding column or row marginals. For the present study, such categories are excluded from investigation, since with zero expected frequencies CFA cannot be carried out. As a result of this consideration, the grade point average variable is treated as an ordinal variable with ten levels and the age at first consumption of alcohol variable is regarded as an ordinal variable with eight levels. Consequently, there are 1,600 configurations in total in this data set. Even for this modified data table, however, many configurations will be empty thereby presenting a challenge to each of the selection procedures presented above.

As for the second data set, there is a total of nine variables, with eight of them being independent and the remaining one dependent. While most of the independent variables have a smaller number of categories than do the ones in the first data set, the challenge here is the number of independent variables is double that of the
former set.

Except for correspondence analysis, all of the above-mentioned preprocessing procedures were carried out using SAS. The PROC LOGISTIC was used to perform the stepwise logistic regression analysis on both of the data sets. The PRU index, also known as the uncertainty coefficient in SAS, which together with the Pearson, the likelihood ratio, and the Mantel-Haenszel chi-square statistics can be assessed by invoking PROC FREQ. A special SAS coding for the CFA method has been adopted with minor modification from that listed in Appendix A of Kennedy (1992). In particular, the BMDPCA procedure was utilized to perform correspondence analysis by virtue of its ease of usage in the present study.

COMPARISON

The five variable selection methods were compared with respect to the order of variables being selected as well as the categories being combined. Special attention was focused on which variable is chosen first by each of the methods. Except for the stepwise logistic regression method, there exists no software precisely for the implementation of any of the selection techniques. Hence, all these strategies are contrasted with regards to their relative ease of execution in terms of the tediousness of the procedures. Also, the reasonableness of the categories
combined will be noted and discussed.

The following table is a summary of all the methods suggested in this methodology chapter pertaining to whether they can handle variable selection or category reduction at the same time.

Table 3. Summary table of the five strategies regarding their capability for handling variable selection and category reduction.

<table>
<thead>
<tr>
<th>Strategies</th>
<th>Variable Selection</th>
<th>Category Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFA</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>PRU</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Pearson chi-square</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Correspondence Analysis</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Stepwise Logistic Regression</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>
CHAPTER IV

RESULTS

FOR THE FIRST DATA SET

A complete cross-classification of all the variables included in the study of freshmen alcohol consumption would constitute a huge table with 1,600 cells, many of which would be empty. The results of each of the data preprocessing techniques on this data set are presented as follows.

CFA approach

Under the first method, i.e. CFA per se, and using frequency of drinking as the dependent variable, there are four one-way combinations, six two-way combinations, four three-way combinations, and one four-way combination with the dependent variable on which configural frequency analyses can be conducted. Even for the one-way combination analyses, there are as many as 12 empty cells out of a total of 50 cells in the g.p.a. by frequency of drink contingency table as well as 11 empty cells out of the 40 for the age by frequency of drink contingency table. Quite expectedly, the
number of empty cells increased rapidly for the three-way combinations and more complicated analyses.

The sum of squares of the $Z$ statistic as well as the corrected $Z$ statistic (denoted $CZ^2$) for CFA are presented in the following table:

Table 4. The sums of squares of the $Z$ and the corrected $Z$ statistics in CFA for the study of freshmen's consumption of alcohol.

<table>
<thead>
<tr>
<th>Variables</th>
<th>N</th>
<th>$\Sigma Z^2$</th>
<th>$\Sigma CZ^2$</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drink x Gender</td>
<td>418</td>
<td>21.07</td>
<td>18.85</td>
<td>7</td>
</tr>
<tr>
<td>Drink x Greek</td>
<td>404</td>
<td>85.80</td>
<td>79.58</td>
<td>7</td>
</tr>
<tr>
<td>Drink x G.P.A.</td>
<td>410</td>
<td>120.54</td>
<td>81.18</td>
<td>47</td>
</tr>
<tr>
<td>Drink x Age</td>
<td>414</td>
<td>162.45</td>
<td>132.03</td>
<td>37</td>
</tr>
<tr>
<td>Drink x Greek x Gender</td>
<td>404</td>
<td>28.05</td>
<td>20.16</td>
<td>16</td>
</tr>
<tr>
<td>Drink x Greek x G.P.A.</td>
<td>397</td>
<td>136.91</td>
<td>84.12</td>
<td>96</td>
</tr>
<tr>
<td>Drink x Greek x Age</td>
<td>400</td>
<td>108.73</td>
<td>73.41</td>
<td>76</td>
</tr>
<tr>
<td>Drink x Gender x G.P.A.</td>
<td>410</td>
<td>45.39</td>
<td>10.97</td>
<td>96</td>
</tr>
<tr>
<td>Drink x Gender x Age</td>
<td>414</td>
<td>32.59</td>
<td>12.65</td>
<td>76</td>
</tr>
<tr>
<td>Drink x G.P.A. x Age</td>
<td>406</td>
<td>602.32</td>
<td>355.41</td>
<td>396</td>
</tr>
<tr>
<td>Drink x Greek x Gender x G.P.A.</td>
<td>397</td>
<td>70.55</td>
<td>18.45</td>
<td>195</td>
</tr>
<tr>
<td>Drink x Greek x Gender x Age</td>
<td>400</td>
<td>54.10</td>
<td>15.88</td>
<td>155</td>
</tr>
<tr>
<td>Drink x Greek x Grade x Age</td>
<td>393</td>
<td>699.60</td>
<td>372.04</td>
<td>795</td>
</tr>
<tr>
<td>Drink x Gender x Grade x Age</td>
<td>406</td>
<td>746.55</td>
<td>369.23</td>
<td>795</td>
</tr>
<tr>
<td>Drink x Greek x Gender x G.P.A. x Age</td>
<td>393</td>
<td>759.66</td>
<td>334.08</td>
<td>1596</td>
</tr>
</tbody>
</table>

Note: An * in the table indicates that some configurations have simultaneous empty row and column marginals.
In the table above, all the combinations marked with an asterisk have some configurations with simultaneous empty row and column marginals such that their expected frequencies are zeroes. Their expected proportions are, likewise, zeroes, thereby creating the indeterminate situation of dividing by zero in the CFA procedure. Under such circumstances, the CFA approach to variable selection is not meaningful. For the sake of completeness, all the cells with zero expected frequency are treated as missing and the corrected $Z^2$ statistic of only those cells with non-zero expected frequency are added up. Hence the sum of the corrected $Z^2$ for three of the three-way combination analyses as well as for all of the four-way combination and the five-way combination analyses should be interpreted with care. Moreover, such treatment has rendered the summated corrected $Z^2$ values of those combinations of variables marked with asterisk as being small, with their corresponding numbers of degrees of freedom. Thus even if the sums of the corrected $Z^2$ of such cells are acceptable, they typically will not be high with respect to their corresponding numbers of degrees of freedom when compared to those of the two-way combination analyses.

In the present study, the analysis of the combination of Greek membership by frequency of drink variable yields the highest sum of corrected $Z^2$ value relative to its number of degrees of freedom. Hence the
Greek variable is the only variable selected and can also be regarded as the first variable selected under the CFA approach.

Further investigation of the CFA of the frequency of drink and Greek affiliation variables reveals that there is one type (at $\alpha^* = .05/10 = .005$), namely, the non-Greek affiliated freshman who has no drinks in a week, and three antitypes, namely, those Greek affiliated freshmen who have no drinks, and the non-Greek freshmen having three and four plus drinks (see Table 5). Hence there is evidence that Greek related freshmen tend to drink more and that the reverse is true for the non-Greek related freshmen.
Table 5. Configural frequency analysis of the frequency of drink and the Greek affiliation variables.

<table>
<thead>
<tr>
<th>Drink-Greek</th>
<th>Observed</th>
<th>Expected Proportion</th>
<th>Corrected Z</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-nG</td>
<td>90</td>
<td>.129</td>
<td>5.548</td>
<td>.00000</td>
</tr>
<tr>
<td>0-G</td>
<td>98</td>
<td>.336</td>
<td>-3.936</td>
<td>.00004</td>
</tr>
<tr>
<td>1-nG</td>
<td>9</td>
<td>.047</td>
<td>-2.256</td>
<td>.01205</td>
</tr>
<tr>
<td>1-G</td>
<td>60</td>
<td>.123</td>
<td>1.456</td>
<td>.07265</td>
</tr>
<tr>
<td>2-nG</td>
<td>6</td>
<td>.030</td>
<td>-1.594</td>
<td>.05549</td>
</tr>
<tr>
<td>2-G</td>
<td>37</td>
<td>.077</td>
<td>1.012</td>
<td>.15574</td>
</tr>
<tr>
<td>3-nG</td>
<td>6</td>
<td>.043</td>
<td>-2.682</td>
<td>.00365</td>
</tr>
<tr>
<td>3-G</td>
<td>57</td>
<td>.113</td>
<td>1.725</td>
<td>.04226</td>
</tr>
<tr>
<td>4-nG</td>
<td>1</td>
<td>.028</td>
<td>-2.968</td>
<td>.00150</td>
</tr>
<tr>
<td>4-G</td>
<td>40</td>
<td>.073</td>
<td>1.882</td>
<td>.02987</td>
</tr>
</tbody>
</table>

Note: The first entry in the Drink-Greek column refers to the number of drinks while G in the second entry denotes belonging to the Greek system and nG denotes not affiliated with the Greek system.

PRU approach

For the second method used, i.e. the PRU index, the results of the forward search procedure are detailed as below. In step 1A, PRU are calculated for each of the two-way tables formed by cross-classifying the dependent variable with each of the independent variables (see Table 6). Judging from the largest PRU indexes, the age of first consumption of alcohol variable (PRU = .138) is regarded as the most important predictor and is therefore selected as the first variable to be included for further study. This first variable is then passed to step 1B, where it is
subjected to an attempt at category reduction. Based on the fact that age is an ordinal variable, the reduction of categories is meaningful only by combining those categories that are adjacent to each other. Of the various feasible reduction undertaken in Table 6, the last entry in step 1B, which indicates the combination of categories 1 with 2, 3 with 4, 5 remaining unchanged, and then 6 and 7 with 9, has resulted in a PRU of .117 and a decrease in $L^2$ by 24.444 ($= 160.154 - 135.710$) and the change in numbers of degrees of freedom is 96 ($= 124 - 28$). Since the change in PRU is not large and the drop in $L^2$ is not significant, the age variable is reduced to four categories in this step.

Table 6. The PRU analysis of the 4 independent variables for the study of freshmen's alcohol consumption.

<table>
<thead>
<tr>
<th>Step 1A</th>
<th>Levels</th>
<th>PRU=$U_{rc}$</th>
<th>$L^2$</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>2</td>
<td>.016</td>
<td>18.643</td>
<td>4</td>
</tr>
<tr>
<td>Greek</td>
<td>2</td>
<td>.069</td>
<td>79.474</td>
<td>4</td>
</tr>
<tr>
<td>G.P.A.</td>
<td>10</td>
<td>.091</td>
<td>105.984</td>
<td>36</td>
</tr>
<tr>
<td>Age</td>
<td>8</td>
<td>.138</td>
<td>160.154</td>
<td>28</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Step 1B Age</th>
<th>Levels</th>
<th>PRU=$U_{rc}$</th>
<th>$L^2$</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>1+2, 3, 4, 5, 6, 7, 9</td>
<td>.126</td>
<td>146.520</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>1, 2+3, 4, 5, 6, 7, 9</td>
<td>.133</td>
<td>155.150</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>1, 2, 3+4, 5, 6, 7, 9</td>
<td>.136</td>
<td>158.418</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>1, 2, 3, 4+5, 6, 7, 9</td>
<td>.121</td>
<td>140.483</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>1, 2, 3, 4, 5+6, 7, 9</td>
<td>.118</td>
<td>137.015</td>
<td>24</td>
<td></td>
</tr>
</tbody>
</table>
In step 2A, the grade point average variable has the largest PRU index (=.213). The corresponding increase

<table>
<thead>
<tr>
<th>Step 2A</th>
<th>Gender</th>
<th>.150</th>
<th>185.25</th>
<th>28</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Greek</td>
<td>.132</td>
<td>163.65</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>G.P.A.</td>
<td>.213</td>
<td>264.14</td>
<td>124</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Step 2B</th>
<th>G.P.A.</th>
<th>.213</th>
<th>264.14</th>
<th>124</th>
</tr>
</thead>
<tbody>
<tr>
<td>0+1, 5, 6, 7, 8, 9, 10, 11, 12</td>
<td>.213</td>
<td>264.14</td>
<td>124</td>
<td></td>
</tr>
<tr>
<td>0+1+5, 6, 7, 8, 9, 10, 11, 12</td>
<td>.211</td>
<td>261.37</td>
<td>120</td>
<td></td>
</tr>
<tr>
<td>0+1+5+6, 7, 8, 9, 10, 11, 12</td>
<td>.204</td>
<td>252.37</td>
<td>108</td>
<td></td>
</tr>
<tr>
<td>0+1+5+6+7, 8, 9, 10, 11, 12</td>
<td>.196</td>
<td>242.27</td>
<td>92</td>
<td></td>
</tr>
<tr>
<td>0+1, 5, 6, 7, 8, 9, 10, 11+12</td>
<td>.203</td>
<td>250.79</td>
<td>108</td>
<td></td>
</tr>
<tr>
<td>0+1+5, 6, 7, 8, 9, 10, 11+12</td>
<td>.200</td>
<td>248.02</td>
<td>104</td>
<td></td>
</tr>
<tr>
<td>0+1+5+6, 7, 8, 9, 10, 11+12</td>
<td>.193</td>
<td>239.02</td>
<td>92</td>
<td></td>
</tr>
<tr>
<td>0+1+5+6+7, 8, 9, 10, 11+12</td>
<td>.185</td>
<td>228.92</td>
<td>76</td>
<td></td>
</tr>
<tr>
<td>0+1+5+6+7, 8, 9, 10+11+12</td>
<td>.177</td>
<td>219.06</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>0+1+5+6+7+8, 9, 10+11+12</td>
<td>.151</td>
<td>187.19</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>0+1+5+6+7+8, 9+10+11+12</td>
<td>.136</td>
<td>168.14</td>
<td>28</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Step 3</th>
<th>Gender</th>
<th>.225</th>
<th>278.21</th>
<th>120</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Greek</td>
<td>.232</td>
<td>286.69</td>
<td>124</td>
</tr>
</tbody>
</table>
in $L^2$ is 128.43 with 116 degrees of freedom. Although the change in $L^2$ is not significant at the .05 level, and since it is only used as a secondary criterion, the grade point average is still chosen based on its relatively large contribution in the PRU. In step 2B, it is found that the decrease in PRU is acceptable if categories 0, 1, 5, 6 and 7 are combined, and categories 10, 11 and 12 are combined. The drop in PRU is .036 and the drop in $L^2$ is 44.08 with 64 degree of freedom. This decrease in $L^2$ is not significant at the .05 level. Hence the grade point average variable can be reduced to an ordinal variable with four levels.

In step 3, both the gender and the Greek variables are being assessed for inclusion as the third independent variable. However, the relative increases of PRU are not large with respect to that in step 2B for both variables. Moreover, the increases in $L^2$ are not significant for either variable. Hence no more variables are selected according to the PRU method.

**Pearson chi-square approach**

According to the Pearson chi-square adjusted for degrees of freedom in the third method, the Greek affiliation variable is judged to have the highest association with the frequency of drink variable ($X^2/df = 18.269$). Hence it is selected as the first variable to enter the model and the search process enters the second
phase. When selecting competing independent variables to be included for further study, the inferential aspect of the Pearson chi-square test is not emphasized, rather it is used as a way to calculate an appropriate index for comparison purposes (see Table 7 on the next page).
Table 7. Variable selection by way of Pearson chi-square per degree of freedom for the study of freshmen's alcohol consumption.

<table>
<thead>
<tr>
<th>Variables</th>
<th>$\chi^2$</th>
<th>df</th>
<th>$\chi^2$/df</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Main effect</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drink x Greek</td>
<td>73.077</td>
<td>4</td>
<td>18.269</td>
</tr>
<tr>
<td>Drink x Gender</td>
<td>18.753</td>
<td>4</td>
<td>4.688</td>
</tr>
<tr>
<td>Drink x G.P.A.</td>
<td>99.198</td>
<td>36</td>
<td>2.756</td>
</tr>
<tr>
<td>Drink x Age</td>
<td>146.186</td>
<td>28</td>
<td>5.221</td>
</tr>
<tr>
<td><strong>1-way interaction</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drink x Greek x Gender</td>
<td>103.628</td>
<td>12</td>
<td>8.636</td>
</tr>
<tr>
<td>Drink x Greek x G.P.A.</td>
<td>160.788</td>
<td>68</td>
<td>2.365</td>
</tr>
<tr>
<td>Drink x Greek x Age</td>
<td>193.909</td>
<td>56</td>
<td>3.463</td>
</tr>
<tr>
<td><strong>2-way interaction</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drink x Greek x Gender x G.P.A.</td>
<td>43.398</td>
<td>28</td>
<td>1.550</td>
</tr>
<tr>
<td>Drink x Greek x Gender x Age</td>
<td>35.879</td>
<td>24</td>
<td>1.495</td>
</tr>
</tbody>
</table>

In the second phase, the Greek variable is combined in turn with each of the remaining independent variables to form one-way interactions. Each of these one-way interactions is cross-classified with the frequency of the drink variable and is then subjected to another Pearson chi-square test. According to the new Pearson chi-square per degree of freedom index, the Greek by gender one-way interaction explains the largest amount of the variance of the dependent variable in comparison with the other interaction terms ($\chi^2$/df = 8.636). In the second stage, the first stopping rule is invoked to determine if the gender variable selected should stay and whether the
procedure should advance to the third stage. For this objective, two separate Pearson chi-square tests are being performed on the gender by frequency of drink contingency table that are adjusted for each of the two levels of the Greek variable. For those freshmen not affiliated with the Greek system, the chi-square statistic is equal to 12.591 with 4 degrees of freedom, whereas for those affiliated with the Greek system, the chi-square amounts to 24.618 with 4 degrees of freedom. Together the chi-square statistics amount to 37.209 with 8 degrees of freedom which is significant at even the 1% significance level. Hence the gender variable is selected as the second variable and the selection procedure moves to the third stage.

In the third stage, two-way interactions are formed by combining the Greek and the gender variables with each of the remaining independent variables. However, the resulting Pearson chi-square per degree of freedom indexes are both very small. Moreover, the termination statistics of the grade point average and the age variable controlling for each level of the Greek by gender one-way interaction are not significant, hence the selection process terminates here with no third variable being selected.

Correspondence analysis approach

In the first round of the correspondence analysis approach, the highest percentage of the total eigenvalue
that is accounted for by the first two axes is 48.5 percent between the frequency of drink and the Greek affiliation variable (see Table 8 on the next page). Hence the Greek variable is chosen as the first variable in the present search. Since it has only two levels, no category reduction is necessary and the selection process enters the second stage.

Before leaving the first round, it has to be pointed out that in both of the plots for the grade point average and the age at first use of alcohol variables, those categories with the fewest frequency counts are always mapped to the corners or the edges of the plots (see the various correspondence plots in Appendix C, especially C3). This phenomenon has rendered the use of proximity in the correspondence plot as a guide in reducing categories impractical, since those categories with low frequency counts, and hence most appropriate for merging, will necessarily be far away from each other.
Table 8. Percentage of total eigenvalue explained by the first two principal axes in the correspondence plot for the study of freshmen's alcohol consumption.

<table>
<thead>
<tr>
<th>Variables</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drink x Greek</td>
<td>48.5</td>
</tr>
<tr>
<td>Drink x Gender</td>
<td>44.2</td>
</tr>
<tr>
<td>Drink x G.P.A.</td>
<td>20.0</td>
</tr>
<tr>
<td>Drink x Age</td>
<td>25.4</td>
</tr>
<tr>
<td>Drink x Greek x Gender</td>
<td>37.4</td>
</tr>
<tr>
<td>Drink x Greek x G.P.A.</td>
<td>16.5</td>
</tr>
<tr>
<td>Drink x Greek x Age</td>
<td>13.3</td>
</tr>
<tr>
<td>Drink x Greek x Gender x G.P.A.</td>
<td>8.7</td>
</tr>
<tr>
<td>Drink x Greek x Gender x Age</td>
<td>9.4</td>
</tr>
</tbody>
</table>

In the second stage, the Greek variable is combined with each of the remaining independent variables to form one-way interaction variables which are then subjected, in turn, to correspondence analyses. It turns out that the first two axes account for 37.4 percent of the eigenvalue in the plot for the one-way interaction between the Greek and gender variables. This value is more than double the size of those for the other two plots for the one-way interaction variables between Greek and g.p.a., and between Greek and age. Hence the gender variable is selected as the second variable to be included.

Entering the third round, the Greek and gender interaction variable is further interacted with the grade
point average and the age variable, respectively, to form two two-way interaction variables. However the percentage of total eigenvalue explained by the first two principal axes is quite low in both cases. Hence the process is discontinued with no third variable being included.

**Stepwise logistic regression approach**

Under the stepwise logistic regression method, all four independent variables are included in the regression equation at the .05 level. The order of entry is, successively, Greek affiliation, grade point average, gender, and age at first consumption of alcohol (see Table 9). Apparently, the stepwise regression approach is the most liberal in comparison to the other methods in the sense that it has selected the greatest number of independent variables.

**Table 9. Order of entry of independent variables as selected by stepwise logistic regression in the study of freshmen's alcohol consumption.**

<table>
<thead>
<tr>
<th>Variables</th>
<th>Order of entry</th>
<th>Score chi-square</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greek</td>
<td>1</td>
<td>66.25</td>
<td>.0001</td>
</tr>
<tr>
<td>Gender</td>
<td>3</td>
<td>18.95</td>
<td>.0001</td>
</tr>
<tr>
<td>G.P.A.</td>
<td>2</td>
<td>34.11</td>
<td>.0001</td>
</tr>
<tr>
<td>Age</td>
<td>4</td>
<td>5.69</td>
<td>.0170</td>
</tr>
</tbody>
</table>
Summary of search

The results of variables selected as well as their order of entry under each method can be conveniently summarized in the following table, Table 10, for ready reference.

Table 10. Summary table of the order of entry with respect to various variable selection strategies.

<table>
<thead>
<tr>
<th>Variables</th>
<th>CFA</th>
<th>PRU</th>
<th>X²/df</th>
<th>CA</th>
<th>Logistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greek</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Gender</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>G.P.A.</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>Age</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>4</td>
</tr>
</tbody>
</table>

From the above table, it is observed that the Greek affiliation variable, which is dichotomous, is chosen as the most important variable in relation to the frequency of drink by most of the selection strategies. In contrast, the age at first consumption of alcohol variable, which has the largest number of categories at 10, is uniquely chosen as the most important variable by the PRU method. While the other four methods are chi-square based methods and favor simple structure (i.e. fewer categories) in the independent variables, the PRU method favors independent variables with more levels and with more uneven distribution of frequencies amongst the levels. Another interesting point is that the gender variable is selected second by Pearson chi-square per degree of freedom and
correspondence analysis but third by stepwise logistic and not at all by the other two methods.

FOR THE SECOND DATA SET

A complete cross-tabulation for all the variables in the teacher educators' productivity study would constitute an extremely large table with 40,320 cells, many of which would be empty. Thus adequate data pre-processing procedures can be very helpful and effective in the understanding of the relationships between the variables involved.

CFA approach

With eight independent variables, there are a total of 8 one-way, 28 two-way, 56 three-way, 70 four-way, 56 five-way, 28 six-way, 28 seven-way and 1 eight-way combinations possible for the CFA approach. Within each of these analyses, the sum of squares of the Z and the corrected Z statistics have to be calculated for all of the possible configurations. Hence this approach is extremely tedious even with the help of computers. Considering the amount of effort in terms of organizing the data preparation for the format specified by the CFA program, and also scrutinizing the results from the first data set, it is determined that CFA will be performed only on the 8
one-way and the 28 two-way combinations. The results of these analyses are summarized in the following Table 11.

Table 11. The sum of square of the Z and the corrected Z statistics in CFA for the study of productivity of teacher educators in Ohio.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Σ Z²</th>
<th>Σ CZ²</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present x Thesis</td>
<td>7.507</td>
<td>6.553</td>
<td>7</td>
</tr>
<tr>
<td>Present x G.P.A.</td>
<td>7.769</td>
<td>5.836</td>
<td>12</td>
</tr>
<tr>
<td>Present x BMAJOR</td>
<td>2.096</td>
<td>1.569</td>
<td>7</td>
</tr>
<tr>
<td>Present x MMAJOR</td>
<td>7.072</td>
<td>5.681</td>
<td>7</td>
</tr>
<tr>
<td>Present x Age</td>
<td>15.296</td>
<td>11.785</td>
<td>17</td>
</tr>
<tr>
<td>Present x Enrolled</td>
<td>11.918</td>
<td>10.384</td>
<td>7</td>
</tr>
<tr>
<td>Present x Teaching</td>
<td>14.697</td>
<td>9.061</td>
<td>27</td>
</tr>
<tr>
<td>Present x Decision</td>
<td>22.163</td>
<td>14.114</td>
<td>32</td>
</tr>
<tr>
<td>Present x Thesis x G.P.A.</td>
<td>28.284</td>
<td>21.794</td>
<td>26</td>
</tr>
<tr>
<td>Present x Thesis x BMAJOR</td>
<td>8.276</td>
<td>5.935</td>
<td>16</td>
</tr>
<tr>
<td>Present x Thesis x MMAJOR</td>
<td>28.847</td>
<td>23.441</td>
<td>16</td>
</tr>
<tr>
<td>Present x Thesis x Age</td>
<td>33.169</td>
<td>23.623</td>
<td>36</td>
</tr>
<tr>
<td>Present x Thesis x Enrolled</td>
<td>16.313</td>
<td>13.381</td>
<td>16</td>
</tr>
<tr>
<td>Present x Thesis x Teaching</td>
<td>38.587</td>
<td>20.598</td>
<td>56</td>
</tr>
<tr>
<td>Present x Thesis x Decision</td>
<td>26.408</td>
<td>19.732</td>
<td>66</td>
</tr>
<tr>
<td>Present x G.P.A. x BMAJOR</td>
<td>17.822</td>
<td>13.440</td>
<td>26</td>
</tr>
<tr>
<td>Present x G.P.A. x MMAJOR</td>
<td>19.315</td>
<td>12.802</td>
<td>26</td>
</tr>
<tr>
<td>Present x G.P.A. x Age</td>
<td>50.481</td>
<td>33.577</td>
<td>56</td>
</tr>
<tr>
<td>Present x G.P.A. x Enrolled</td>
<td>22.927</td>
<td>16.665</td>
<td>26</td>
</tr>
<tr>
<td>Present x G.P.A. x Teaching</td>
<td>77.347</td>
<td>42.868</td>
<td>86</td>
</tr>
<tr>
<td>Present x G.P.A. x Decision</td>
<td>53.643</td>
<td>40.500</td>
<td>101</td>
</tr>
<tr>
<td>Present x BMAJOR x MMAJOR</td>
<td>65.163</td>
<td>57.862</td>
<td>16</td>
</tr>
<tr>
<td>Present x BMAJOR x Age</td>
<td>31.770</td>
<td>22.755</td>
<td>36</td>
</tr>
</tbody>
</table>
Inspection of the table above reveals that the two-way combination of the bachelor's major and the master's major variables with the dependent variable has the largest value for the sum of squares of the corrected Z statistic with respect to its number of degrees of freedom. A follow-up study by configural frequency analysis for this particular combination of variables is presented as Table 12 on the following page.
Table 12. Configural frequency analysis of the number of presentation, bachelor's major in education, and master's major in education variables.

<table>
<thead>
<tr>
<th>Cell</th>
<th>Observed</th>
<th>Residual</th>
<th>CZ</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0YY</td>
<td>125</td>
<td>11.6400</td>
<td>1.17052</td>
<td>0.12090</td>
</tr>
<tr>
<td>0YN</td>
<td>23</td>
<td>-11.6380</td>
<td>-1.95341</td>
<td>0.02539</td>
</tr>
<tr>
<td>1YY</td>
<td>35</td>
<td>5.1280</td>
<td>0.87012</td>
<td>0.19212</td>
</tr>
<tr>
<td>1YN</td>
<td>4</td>
<td>-5.1277</td>
<td>-1.54429</td>
<td>0.06126</td>
</tr>
<tr>
<td>2YY</td>
<td>31</td>
<td>4.1910</td>
<td>0.73043</td>
<td>0.23256</td>
</tr>
<tr>
<td>2YN</td>
<td>4</td>
<td>-4.1915</td>
<td>-1.29927</td>
<td>0.09693</td>
</tr>
<tr>
<td>3YY</td>
<td>20</td>
<td>2.3830</td>
<td>0.45580</td>
<td>0.32427</td>
</tr>
<tr>
<td>3YN</td>
<td>3</td>
<td>-2.3830</td>
<td>-0.81549</td>
<td>0.20739</td>
</tr>
<tr>
<td>4YY</td>
<td>68</td>
<td>12.0850</td>
<td>1.63231</td>
<td>0.05131</td>
</tr>
<tr>
<td>4YN</td>
<td>5</td>
<td>-12.0850</td>
<td>-2.84622</td>
<td>0.00221</td>
</tr>
<tr>
<td>ONY</td>
<td>70</td>
<td>-21.9150</td>
<td>-2.44149</td>
<td>0.00731</td>
</tr>
<tr>
<td>ONN</td>
<td>50</td>
<td>21.9150</td>
<td>4.14546</td>
<td>0.00002</td>
</tr>
<tr>
<td>1NY</td>
<td>25</td>
<td>1.2550</td>
<td>0.15831</td>
<td>0.43711</td>
</tr>
<tr>
<td>1NN</td>
<td>6</td>
<td>-1.2553</td>
<td>-0.28223</td>
<td>0.38888</td>
</tr>
<tr>
<td>2NY</td>
<td>18</td>
<td>-4.2130</td>
<td>-0.80380</td>
<td>0.21076</td>
</tr>
<tr>
<td>2NN</td>
<td>11</td>
<td>4.2128</td>
<td>1.43379</td>
<td>0.07582</td>
</tr>
<tr>
<td>3NY</td>
<td>14</td>
<td>-1.3190</td>
<td>-0.21215</td>
<td>0.41599</td>
</tr>
<tr>
<td>3NN</td>
<td>6</td>
<td>1.3191</td>
<td>0.38017</td>
<td>0.35191</td>
</tr>
<tr>
<td>4NY</td>
<td>26</td>
<td>-9.2340</td>
<td>-1.51964</td>
<td>0.06430</td>
</tr>
<tr>
<td>4NN</td>
<td>20</td>
<td>9.2340</td>
<td>2.68764</td>
<td>0.00360</td>
</tr>
</tbody>
</table>

Note: The first entry in the cells of the first column refers to the number of presentation at conferences, and the Y in the second and the third entry refers to teacher educators who are education majors at the bachelor and master level respectively, while the N refers to non-education majors for those two levels. CZ in the fourth column refers to the corrected Z statistic.

Closer scrutiny of the results disclosed that the
large corrected Z value mainly stems from a significant type composed of 50 students who had not majoring in education at the bachelor and the masters levels who had never presented at any conferences. The p-value for the corrected unit normal approximation to the binomial distribution for this configuration is .00002. Hence the two independent variables are selected as important variables in the CFA approach. Since they are chosen simultaneously, the order of entry issue is not relevant for this method.

**PRU approach**

From Table 13 it can be seen that the decision variable has the highest PRU index in the first round and is hence chosen as the first important variable. Subsequent category reduction attempts indicate that by combining levels 1, 2 with 3, 4 with 5, and 6 with 7 yields a PRU of .0067. This reflects a drop of .0062 in the PRU. Correspondingly, there is a drop by 6.47 in $L^2$ while the number of degrees of freedom is decreased by 16. This change in $L^2$ is not significant at the .05 level. Guided by the principle of parsimony, the categories of the decision variable is reduced into a variable with 3 levels (i.e. 1+2+3, 4+5, 6+7) as it enters the second phase of the search.

All the PRU indexes of the independent variables
increase substantially in the second phase, with the teaching variable having the largest value at 0.0354. The increase in $L^2$ is 45.14 (\(55.68 - 10.54\)) and the corresponding increase in the number of degrees of freedom is 60 (\(68 - 8\)). This increase is not significant even at the .10 level. Thus based on the termination rule, the search is stopped.
Table 13. Using PRU as a criterion in selecting variables and reducing categories for the study of productivity of teacher educators in Ohio.

<table>
<thead>
<tr>
<th>Step 1A</th>
<th>Levels</th>
<th>PRU(=U_{ijc})</th>
<th>(L^2)</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thesis</td>
<td>2</td>
<td>.0041</td>
<td>6.50</td>
<td>4</td>
</tr>
<tr>
<td>G.P.A.</td>
<td>3</td>
<td>.0052</td>
<td>8.14</td>
<td>8</td>
</tr>
<tr>
<td>BMAJOR</td>
<td>2</td>
<td>.0012</td>
<td>1.81</td>
<td>4</td>
</tr>
<tr>
<td>MMAJOR</td>
<td>2</td>
<td>.0040</td>
<td>6.29</td>
<td>4</td>
</tr>
<tr>
<td>Age</td>
<td>4</td>
<td>.0101</td>
<td>15.80</td>
<td>12</td>
</tr>
<tr>
<td>Enrolled</td>
<td>2</td>
<td>.0067</td>
<td>10.55</td>
<td>4</td>
</tr>
<tr>
<td>Teaching</td>
<td>6</td>
<td>.0096</td>
<td>15.07</td>
<td>20</td>
</tr>
<tr>
<td>Decision</td>
<td>7</td>
<td>.0139</td>
<td>21.90</td>
<td>24</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Step 1B Decision</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1+2, 3, 4, 5, 6, 7</td>
<td>.0107</td>
<td>16.79</td>
<td>20</td>
</tr>
<tr>
<td>1, 2+3, 4, 5, 6, 7</td>
<td>.0088</td>
<td>16.82</td>
<td>20</td>
</tr>
<tr>
<td>1, 2, 3+4, 5, 6, 7</td>
<td>.0126</td>
<td>19.82</td>
<td>20</td>
</tr>
<tr>
<td>1, 2, 3, 4+5, 6, 7</td>
<td>.0135</td>
<td>21.21</td>
<td>20</td>
</tr>
<tr>
<td>1, 2, 3, 4, 5+6, 7</td>
<td>.0092</td>
<td>14.51</td>
<td>20</td>
</tr>
<tr>
<td>1, 2, 3, 4, 5+6+7</td>
<td>.0119</td>
<td>18.69</td>
<td>20</td>
</tr>
<tr>
<td>1+2+3, 4+5, 6+7</td>
<td>.0067</td>
<td>10.54</td>
<td>8</td>
</tr>
<tr>
<td>1+2+3+, 4, 5, 6, 7</td>
<td>.0092</td>
<td>14.43</td>
<td>16</td>
</tr>
<tr>
<td>1+2+3, 4+5, 6, 7</td>
<td>.0087</td>
<td>13.74</td>
<td>12</td>
</tr>
<tr>
<td>1+2+3, 4, 5, 6+7</td>
<td>.0071</td>
<td>11.22</td>
<td>12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Step 2</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Thesis</td>
<td>.0169</td>
<td>26.56</td>
<td>20</td>
</tr>
<tr>
<td>G.P.A.</td>
<td>.0281</td>
<td>44.24</td>
<td>32</td>
</tr>
<tr>
<td>BMAJOR</td>
<td>.0137</td>
<td>21.52</td>
<td>20</td>
</tr>
<tr>
<td>MMAJOR</td>
<td>.0189</td>
<td>29.75</td>
<td>20</td>
</tr>
<tr>
<td>Age</td>
<td>.0276</td>
<td>43.41</td>
<td>44</td>
</tr>
<tr>
<td>Enrolled</td>
<td>.0168</td>
<td>26.36</td>
<td>20</td>
</tr>
<tr>
<td>Teaching</td>
<td>.0354</td>
<td>55.68</td>
<td>68</td>
</tr>
</tbody>
</table>
Pearson chi-square approach

According to Table 14, the first variable selected is the enrollment variable, which has the largest Pearson chi-square per degree of freedom (= 2.583). This variable is then combined with each of the remaining independent variables, and cross-tabulated with the presentation variable. It is found that the enrollment by major at master's level has the largest Pearson chi-square per degree of freedom (= 1.905). However, the subsequent assessment of the unique contribution of master's major variable, controlling for the enrollment variable already selected, generates a chi-square equal to 13.601 (= 6.138 + 7.463) at 8 (= 4 + 4) degrees of freedom, which is not significant at the .05 level. Moreover, the second terminating statistic, the extended Mantel-Haenszel chi-square is also not significant (= 7.009 at 4 degrees of freedom) at the .05 level. Hence the master's major variable falls short of being selected as the second variable by this method.
Table 14. Variables selection by way of Pearson chi-square adjusting for the degree of freedom for the study of productivity of teacher educators.

<table>
<thead>
<tr>
<th>Variables</th>
<th>X²</th>
<th>df</th>
<th>X²/df</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Main effect</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Present x Thesis</td>
<td>6.486</td>
<td>4</td>
<td>1.622</td>
</tr>
<tr>
<td>Present x G.P.A.</td>
<td>7.353</td>
<td>8</td>
<td>0.919</td>
</tr>
<tr>
<td>Present x BMAJOR</td>
<td>1.800</td>
<td>4</td>
<td>0.450</td>
</tr>
<tr>
<td>Present x MMAJOR</td>
<td>5.973</td>
<td>4</td>
<td>1.493</td>
</tr>
<tr>
<td>Present x Age</td>
<td>14.489</td>
<td>12</td>
<td>1.207</td>
</tr>
<tr>
<td>Present x Enrolled</td>
<td>10.333</td>
<td>4</td>
<td>2.583</td>
</tr>
<tr>
<td>Present x Teaching</td>
<td>13.066</td>
<td>20</td>
<td>0.653</td>
</tr>
<tr>
<td>Present x Decision</td>
<td>19.614</td>
<td>24</td>
<td>0.817</td>
</tr>
<tr>
<td><strong>1-way interaction</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Present x Enrolled x Thesis</td>
<td>18.674</td>
<td>12</td>
<td>1.556</td>
</tr>
<tr>
<td>Present x Enrolled x G.P.A.</td>
<td>29.086</td>
<td>20</td>
<td>1.455</td>
</tr>
<tr>
<td>Present x Enrolled x BMAJOR</td>
<td>18.761</td>
<td>12</td>
<td>1.563</td>
</tr>
<tr>
<td>Present x Enrolled x MMAJOR</td>
<td>22.862</td>
<td>12</td>
<td>1.905</td>
</tr>
<tr>
<td>Present x Enrolled x Age</td>
<td>31.775</td>
<td>28</td>
<td>1.135</td>
</tr>
<tr>
<td>Present x Enrolled x Teaching</td>
<td>39.948</td>
<td>44</td>
<td>0.908</td>
</tr>
<tr>
<td>Present x Enrolled x Decision</td>
<td>52.948</td>
<td>52</td>
<td>1.013</td>
</tr>
</tbody>
</table>

**Correspondence analysis approach**

Note that for those independent variables with few categories, the first two principal axes of their correspondence plots tend to explain higher percentages of the total eigenvalues involved (see Table 15). Moreover, their percentages are quite close to each other. The highest one which is associated with the plot on presentation and the enrollment variables is only 0.6% in
excess of the second highest percentage. Since enrollment
has only two levels, no combination of categories is
necessary here.

Table 15. Percentage of total eigenvalue explained by the
first two principal axes in correspondence plot
for the study of productivity.

<table>
<thead>
<tr>
<th>Variables</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present x Thesis</td>
<td>42.1</td>
</tr>
<tr>
<td>Present x G.P.A.</td>
<td>36.2</td>
</tr>
<tr>
<td>Present x BMAJOR</td>
<td>41.1</td>
</tr>
<tr>
<td>Present x MMAJOR</td>
<td>42.1</td>
</tr>
<tr>
<td>Present x Age</td>
<td>31.7</td>
</tr>
<tr>
<td>Present x Enrolled</td>
<td>42.7</td>
</tr>
<tr>
<td>Present x Teaching</td>
<td>24.4</td>
</tr>
<tr>
<td>Present x Decision</td>
<td>22.2</td>
</tr>
<tr>
<td>Present x Enrolled x Thesis</td>
<td>31.9</td>
</tr>
<tr>
<td>Present x Enrolled x G.P.A.</td>
<td>25.9</td>
</tr>
<tr>
<td>Present x Enrolled x BMAJOR</td>
<td>32.1</td>
</tr>
<tr>
<td>Present x Enrolled x MMAJOR</td>
<td>32.6</td>
</tr>
<tr>
<td>Present x Enrolled x Age</td>
<td>21.2</td>
</tr>
<tr>
<td>Present x Enrolled x Teaching</td>
<td>15.5</td>
</tr>
<tr>
<td>Present x Enrolled x Decision</td>
<td>13.9</td>
</tr>
<tr>
<td>Present x Enrolled x MMAJOR x Thesis</td>
<td>21.2</td>
</tr>
<tr>
<td>Present x Enrolled x MMAJOR x G.P.A.</td>
<td>16.1</td>
</tr>
<tr>
<td>Present x Enrolled x MMAJOR x BMAJOR</td>
<td>21.0</td>
</tr>
<tr>
<td>Present x Enrolled x MMAJOR x Age</td>
<td>12.9</td>
</tr>
<tr>
<td>Present x Enrolled x MMAJOR x Teaching</td>
<td>8.3</td>
</tr>
<tr>
<td>Present x Enrolled x MMAJOR x Decision</td>
<td>8.3</td>
</tr>
</tbody>
</table>
In the second round, the first two axes associated with the master's major variable accounts for 32.6% of the total eigenvalue and is thus selected as the second important variable. However, the four levels of the enrollment by master's major variable defy any meaningful category reduction and the reduction is not attempted.

The percentages for the three-way combinations are judged to be quite low and the process is subjectively terminated at this point.

**Stepwise logistic regression approach**

According to the stepwise logistic regression method, out of the eight presumably meaningful independent variables considered, only three are included in the regression equation at the .05 level. The order of entry is, sequentially, whether the educators were enrolled as full-time students, whether they had written master theses and whether they were education majors in graduate study (see Table 16). Apparently, the stepwise regression approach is the most liberal in comparison to the other methods in the sense that it has selected the greatest number of independent variables.
Table 16. Order of entry of independent variables selected by stepwise logistic regression for the study of productivity of teacher educators.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Order of entry</th>
<th>Score chi-square</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enrolled</td>
<td>1</td>
<td>7.6283</td>
<td>.0057</td>
</tr>
<tr>
<td>Thesis</td>
<td>2</td>
<td>5.2157</td>
<td>.0224</td>
</tr>
<tr>
<td>MMAJOR</td>
<td>3</td>
<td>4.5222</td>
<td>.0335</td>
</tr>
</tbody>
</table>

Summary of search

The results of variables being selected by as well as their order of entry under each method can be conveniently summarized in Table 17 as follows for ready reference.

Table 17. Summary table of order of entry of variables with respect to the various variable selection strategies.

<table>
<thead>
<tr>
<th>Variables</th>
<th>CFA</th>
<th>PRU</th>
<th>X²/df</th>
<th>CA</th>
<th>Logistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thesis</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>G.P.A.</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BMAJOR</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MMAJOR</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Age</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Enrolled</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Teaching</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Decision</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

The ticks (✓) in the table above denote that the
variables are selected at the same time with no reference to order. From the above table, it is observed that the enrollment status variable, which is dichotomous, is chosen as the most important variable in relation to the frequency of conference presentation by the chi-square per degree of freedom, correspondence analysis, and the logistic regression methods. The major at master's degree variable, which is also dichotomous, is chosen by the CFA, the correspondence and the logistic approaches. In contrast, the decision to become a teacher educator variable, which has the greatest number of categories at 7, is uniquely chosen as the most important variable by the PRU method. Hence again it is observed that the chi-square based methods favor simple structure (i.e. fewer categories) in the independent variables, whereas the PRU method favors independent variables with more levels and with more uneven distributions of frequencies amongst the levels.
CHAPTER V

DISCUSSION

GENERAL COMMENTS

Based on the results from applying the five different strategies to the two data sets used in the present empirical and exploratory study, some general impressions regarding this set of strategies are first presented before considering each of the methods individually.

The first general impression is that even though each of the methods uses statistical calculations to guide the process in the selection of variables and/or the reduction of categories, much subjectivity is called for within each of the methods except in the case of the CFA approach. Hence these methods are as much an art as they are objective endeavor.

It should also be pointed out that, except for the CFA per se approach, all of the strategies used in the present study are stepwise in nature. In addition, they are all formulated in a forward fashion. One of the major
disadvantages with stepwise searching procedures is that they only produce a reasonable subsets of variables according to the criteria for those particular methods. The difficulty hinges on the fact that stepwise procedures consider variables one at a time, rather than making use of the joint information among several variables at the same time. Hence the results are not fool-proof and the subsets generated may be very different from method to method. Consequently, it is not applicable to use the term "the best subset" resulting from a particular stepwise procedure.

Another difficulty with stepwise procedures is that their exhaustive testing feature is essentially capitalizing on chance. Moreover, the stepwise searching process is usually very tedious unless tailored-made software programs are written for this specific purpose.

From a practical point of view, since different subsets of variables may be selected by different methods, different interpretations are possible depending on which variable selection approach is adopted. This would create ambiguity as far as the purpose of substantive understanding of the research problem is concerned.

According to the literature on stepwise methods for such parametric techniques as multiple regression and discriminant analysis (see, e.g., Huberty, 1989), one of the most common criticisms about the stepwise approach is
in its automatic searching feature that defies the control of the researcher. This issue is not of major concern in the present study, since except for the CFA and the stepwise logistic regression approaches, all the other methods allow for much maneuverability on the part of the researcher to apply his or her own judgment. In fact, in the case of correspondence analysis, the freedom that is permissible to the interpretation of the correspondence plots allow much subjectivity to the researcher.

One last comment is that in order to do justice to the various variable selection techniques in terms of comparing their performances, some category reduction consideration for the data set should be conducted prior to the application of any variable selection procedures. Rather than relying on intuition as a guideline, which is the common practice, there is a need for some objective criteria to help practitioners make their judgments. Since part of the purpose of the present study is to compare the relative merits between the PRU index and the correspondence analysis approach in reducing categories of variables in the same data set, the aforementioned data pre-processing treatment is not being conducted.

Discussion of the CFA approach

The CFA approach is extremely tedious when the number of variables of interest is large. Practicality
will be a big concern in this circumstance. Even for a data set of moderate size, the expected frequencies of some of the cells may be equal to zero and thus create technical difficulty. This difficulty, however, is not specific to CFA, since other methods that involve expected frequencies in their calculations will encounter the same problem. Besides, there is the theoretical weakness that the configurations in a typical CFA are not independent of each other. The use of the binomial distribution will not do away with this fact, and is rather inappropriate in the first place. Despite all these weaknesses, if the purpose of the study is to find out if a group of subjects with particular characteristics appear either more or less often than by chance, then CFA may shed some light along this direction. The CFA results from the two data sets in the present study illustrate the existence of groups of subjects with particular features quite well. Still, overall speaking, CFA must be used with much caution as an analytical or a variable selection technique.

Discussion of the PRU approach

Clark et al. (1986, 1988) indicate that, according to their experiences, there is evidence that the PRU approach performs better than the Pearson chi-square per degree of freedom method suggested by Higgins and Koch (1977). On the one hand, the PRU approach can generate a
subset of variables that has high explanatory power for the variation of the dependent variable. On the other hand, the Pearson chi-square method emphasizes the relative importance of relations in comparison with one another rather than degree of association as in the PRU index. Hence "the distinction is comparable with that between analysis of quantitative data with the help of the squared correlation coefficient (which indicates proportional reductions in variance, that is, degree of correlation) and an analysis based on absolute reductions in variance, that is, using only the numerator of this coefficient (1986, p.769)".

The PRU index and the correspondence analysis approach can handle both variable selection and category reduction at the same time. The other three methods can only be used to select variables. Even though category reduction can be guided by referring to the PRU index, subjective judgment on the part of the researcher is still necessary. The interplay of the PRU index and the termination statistic as suggested in the literature can be very subjective as well.

As it is pointed out earlier, PRU favors, in general, variables with more categories while the Pearson chi-square approach favors variables with fewer categories. In other words, PRU favors a more complicated structure while chi-square favors simple structure. Since the CFA
and the stepwise logistic approach are basically dependent on some form of the chi-square statistic, they all favor variables with fewer categories, in every round of the search.

**Discussion of the Pearson chi-square approach**

The Pearson chi-square approach is similar to the stepwise logistic regression approach in the first round. They differ, however, starting from the second round, since the Pearson chi-square takes into consideration the interaction effect on the dependent variable while the stepwise logistic regression does not. The stopping rules used in this approach, however, depend more on the main effects while controlling for the effects of those variables already chosen. Computationally speaking, this procedure is quite easy to perform in comparison to the other methods examined in the present study.

**Discussion of the correspondence analysis approach**

The use of correspondence analysis as a guide to reducing categories does not perform as well as had been expected. For example, in the study of freshmen's consumption of alcohol, it is noticed that those categories with the fewest frequency counts (say, the category 26+ which has only one cell count in the age variable in Appendix B4), are always mapped to the corners or the edges
of the plots. This phenomenon can be explained by the fact that such categories are quite different from the average row or column profile, or centroids, which is the average representation of all the row or column categories. This average profile is usually being used to anchor the correspondence plot and hence serves as the origin of the plot (see Lebart et al., 1984). As a result, such extreme categories will have little chance of falling in the proximity of other categories to be considered for combination.

The subsets of variables selected by means of correspondence analysis parallel closely those selected by the Pearson chi-square method. This is due to the fact that the total eigenvalue in correspondence analysis is equivalent to Pearson chi-square divided by the total sample size (see Appendix B). As a result, the percentage of total eigenvalue explained by the first two principal axes is related to the Pearson chi-square, therefore the two apparently different criteria will yield similar results.

Discussion of stepwise logistic regression

The results from stepwise logistic regression are the most liberal of all, in the sense that it selects more variables as important variables than any of the other methods adopted in the present study. One possible reason
is that all of the other methods include consideration of interaction effects among the independent variables while the stepwise logistic regression considers only the main effects terms. Thus it is advisable to regard the necessity of including the interaction terms before the execution of a stepwise logistic regression.

LIMITATIONS OF STUDY

As has been pointed out above, all the methods utilized in this study, except CFA, are stepwise procedures. As a consequence they all carry the same disadvantage of stepwise approaches, that is, the outcome group of variables only represents a reasonable subset according to certain criteria. They may or may not constitute the best subset of variables for further study.

Moreover, the mode of inquiry employed in this study is asymmetrical in nature. Hence the observation and discussion above may or may not be applicable to studies in which there exist no independent and dependent variable framework. In other words, when the purpose is to search for a subset of variables that are associated with each other and to reduce the numbers of categories of the variables, the above discussion may or may not be extendable to these other types of studies.

On the other hand, owing to the exploratory nature of the present study, the literature on collapsibility of
contingency tables, which is highly technical, has not been referred to for guidelines in the reduction of categories.

In carrying out the CFA approach in the present study, no special consideration has been made to the existence of the issue of sampling zeroes. Some of the expected frequencies under this situation may be zero and thus create the problem discussed above. This issue needs to be tackled by further study.

RECOMMENDATIONS

Despite the limitations mentioned above, and based on the experience gained in the present endeavor, some guidelines can still be made regarding what techniques should be used to select variables.

The first suggestion is to choose variables that are important on theoretical grounds and force them into the selection process in the first place. However, for the sake of parsimonious and stability, only variables that are really important, as evidenced from the literature, should be included.

If only one method is to be used, then it is suggested to use PRU as the single method for both variable selection and category reduction. However, it may be better to use a combination of approaches, and at the same time be guided by the purpose of the study. For example, for a study with few variables, if the purpose is to find
clusters of subjects with distinct characteristics, then the method of choice may be the CFA approach.

Also it is suggested that category reduction should first be applied to the variables of interest before any variable selection takes place. The PRU index is a good method of choice in this regard.

Finally, it is suggested that the validity of the subset of variables selected by any process should be confirmed by a cross-validation study.

FUTURE STUDY

Since categorical variable selection and category reduction are important methodological considerations, and since the methods explored in the present study have their own weaknesses, more research is necessary in this area. There are several avenues to pursue such an endeavor.

First, the effects of the above methods on symmetrical studies should also be studied. Intuitively speaking, these strategies can still be applicable by considering, in turn, each variable as the dependent variable and carry out the mechanism of each method. However, such endeavors can be extremely tedious for studies with numerous variables. On the other hand, there may exist more appropriate methods that cater towards this kind of inquiry.

Another meaningful direction is to conduct
simulation studies by applying the various categorical variable selection methods on data that have known structure or properties built into it. The relative performance of the different variable selection methods with respect to the ease and extent of recovery of information of the data structure can then be used as indicators of the effectiveness of these techniques. More research need to be done in this area.

In recent years, fuzzy set theory is gaining the interest of a lot of researchers in various disciplines. There is a growing body of literature on its applications. There has been some works done on defining some fuzzy likeness coefficients as a way to measure the degree of resemblance between two rows or columns in the case of categorical data. Such coefficients can be compared to the performance of the PRU index as well as other similarity indices as indicators of eligibility for combining categories.

Lebart et al. (1984) caution against using the percentages of total eigenvalue as a way to characterize the quality of a correspondence analysis based on their experience. They explain that they have encountered counter-examples in which the percentages of explained eigenvalue represents only a conservative idea of the amount of information being presented by a correspondence plot. More research should be devoted to this issue,
especially because correspondence analysis is highly recommended by the French school as a form of exploratory data analysis for categorical data.

More appropriate to the case of correspondence analysis as a guidepost for combining categories is to use the chi-square metric that is intrinsic to correspondence analysis to calculate the distances between a pair of rows or of columns. This has actually been suggested by Orton and Tyers (1990) as a way to reduce sparse contingency tables rather than basing the decision entirely on subjective judgment. However, they admitted that there are, at times, meaningless merges. Moreover, they also encounter some difficulty for rows or columns that have very few categories. Usually their distances are far away from other rows and columns and thus defy the purpose of using the chi-square metric as a guide to combine categories. Again, improvement of the use of this metric may be a valuable avenue to pursue. In fact, Orton and Tyers (1990) mentioned that they are working on some simulation studies on this metric in order to understand more of its performance.

Finally, there is a growing body of literature on collapsibility. However, these studies are highly technical at the present moment. More work need to be done to bring these techniques into terms understandable by researchers for practical uses. Their comparisons with the
methods employed in the present study might be very interesting.
APPENDIX A

TAXONOMY OF MODELS FOR QUALITATIVE DATA
APPENDIX A

Taxonomy of Models for Qualitative Data

<table>
<thead>
<tr>
<th>Model</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Latent Attribute Models</td>
<td></td>
</tr>
<tr>
<td>a. Latent structure models</td>
<td>Larzarsfeld &amp; Henry (1968)</td>
</tr>
<tr>
<td>Latent class model</td>
<td>Goodman (1974)</td>
</tr>
<tr>
<td>Latent class with response error</td>
<td>Dayton &amp; MacReady (1980)</td>
</tr>
<tr>
<td>b. Scalability models</td>
<td></td>
</tr>
<tr>
<td>Scalogram analysis</td>
<td>Guttman (1950)</td>
</tr>
<tr>
<td>Probabilistic scale analysis</td>
<td>Proctor (1970)</td>
</tr>
<tr>
<td>Multiple scale analysis</td>
<td>Mokken (1971)</td>
</tr>
<tr>
<td>Bi- &amp; multiform scales</td>
<td>Goodman (1975a, 1975b)</td>
</tr>
<tr>
<td>Probabilistic validation</td>
<td>Dayton &amp; MacReady (1976)</td>
</tr>
<tr>
<td>Order analysis</td>
<td>Krus (1976)</td>
</tr>
<tr>
<td>Scaling of order hypothesis</td>
<td>Davison (1979, 1980)</td>
</tr>
<tr>
<td>Probabilistic unfolding</td>
<td>Coombs &amp; Smith (1973)</td>
</tr>
<tr>
<td>Quasi-independence model</td>
<td>Goodman (1975)</td>
</tr>
<tr>
<td>c. Latent trait models</td>
<td></td>
</tr>
<tr>
<td>Rasch model</td>
<td>Rasch (1966)</td>
</tr>
<tr>
<td>Normal ogive model</td>
<td>Lord &amp; Novick (1968)</td>
</tr>
<tr>
<td>Three-parameter logistic model</td>
<td>Birnbaum (1968)</td>
</tr>
<tr>
<td>Logistic change model</td>
<td>Fischer (1976)</td>
</tr>
<tr>
<td>d. Factor analysis models</td>
<td></td>
</tr>
<tr>
<td>FA for dichotomous variables</td>
<td>Muthén (1978); Christofferson (1975)</td>
</tr>
<tr>
<td>Monotonicity analysis</td>
<td>Bentler (1970)</td>
</tr>
</tbody>
</table>

1 What Henning and Rudinger refer to as "qualitative" data are equivalent to "categorical" or "discrete" data as referred to by other researchers.
2. Prediction Models

a. Dichotomous regression model
b. Structural equation models
   LISREL
   Jöreskog & Sörbom (1981)
   Partial least squares
   Wold (1979)
c. Cross-classification with error models
   Prediction analysis
   Hildebrand et. al. (1977)
   Fitting cross-classification
   Thomas (1977)
   Matching model
   Hubert (1979)
d. Multidimensional contingency tables with partial least squares
   Wold & Bertholet (1983)

3. Multinomial Response Models
a. Log-linear model
b. Correspondence analysis
c. Dual scaling for categorical data

Note: 1 This taxonomy was initiated by Bentler (1980) and subsequently expanded by Henning and Rudinger (1985).
2 Here log-linear model is regarded as an umbrella term that incorporates both the symmetrical mode of inquiry as well as the asymmetrical mode, which is also commonly known as logit analysis.
APPENDIX B
AN INTRODUCTION TO CORRESPONDENCE ANALYSIS
APPENDIX B

An Introduction to Correspondence Analysis

This appendix provides a succinct introduction to correspondence analysis (CA). The notation used here matches that used for the schematic diagram of CA as presented in Figure 2 in Chapter Three. The mechanism demonstrated there is CA represented as a form of principal component analysis (PCA). It is formulated after the fashion presented in Pielou (1984). The mathematical derivation introduced here, however, is a slight variation and is compiled primarily with reference to the formulation as presented in Leclerc et al. (1988) and Everitt (1988).

Define

\[ f_{i.} = \sum_j f_{ij}, \quad p_{i.} = \frac{f_{i.}}{N} \]
\[ f_{.j} = \sum_i f_{ij}, \quad p_{.j} = \frac{f_{.j}}{N} \]

where \( i = 1, 2, \ldots, r \) and \( j = 1, 2, \ldots, c \).

Clearly,
Let \( N = \sum_i f_{i*} = \sum_j f_{.j} = \sum_i \sum_j f_{ij} \).

Let

\[ d_{ij} = \frac{f_{ij} - N \pi_{i} \pi_{j}}{\sqrt{N \pi_{i} \pi_{j}}} \]

where \( N \pi_{i} \pi_{j} \) is the expected frequency of cell \((i,j)\) under the independence model. Further, let \( D \) be a \( r \times c \) matrix with \( d_{ij} \) as the element in the \( i \)th row and \( j \)th column. An application of the singular value decomposition (SVD), also known as finding the basic structure in social sciences, of \( D \) will then result in a product of three matrices, \( U \), \( V \), and \( \Lambda \), such that

\[ D = U \Sigma V' \]

where \( U \) is an orthonormal matrix with \( UU' = I \), \( V \) is an orthogonal matrix with \( VV' = V'V = I \), and \( \Lambda \) is a diagonal matrix with non-negative diagonal elements, known as singular values, arranged in descending order of magnitude.

By extracting a factor of \( N^{1/2} \) from the right hand side of the decomposition above, the matrix \( D \) can be re-expressed as follows:

\[ D = \sqrt{N} U' \Lambda' V' \]

Without loss of generality, the asterisks in the matrices can be dropped for simplicity in the following derivation.

Since \( D' = \sqrt{N} V^\prime U' \), therefore \( DD' = N U \Lambda^2 U' \) and \( D'D = N V \Lambda^2 V' \). Thus \( U \) contains the eigenvectors of \( 1/N DD' \),
while $V$ contains those of $1/N \ D' \ D$. Notice also that the square of the singular values on the diagonal of $A^2$ are the same as the eigenvalues of $1/N \ DD'$, or, equivalently, $1/N \ D' \ D$.

Let $R$ be the rank of matrix $D$. Then each element of $D$, based on the SVD, can be represented by the expression below:

$$d_{ij} = \sum_{k=1}^{R} \sqrt{\lambda_k} u_{ik} v_{jk},$$

where $u_{ik}$ and $v_{jk}$ are the elements in the $k$th column of matrices $U$ and $V$, respectively, while $\lambda_k$ is the $k$th eigenvalue of the matrix $1/N \ DD'$.

The sum of the eigenvalues of $1/N \ DD'$, which is also known as inertia in the CA literature, can be found by taking the trace of the matrix $1/N \ DD'$:

$$\text{tr}(\frac{1}{N} DD') = \sum_k \lambda_k$$
$$= \frac{1}{N} \text{tr}(DD')$$
$$= \frac{1}{N} \sum_i \sum_j d_{ij}^2$$
$$= \frac{X^2}{N}.$$

Here, the $X^2$ is Pearson's chi-square statistic for the $r \times c$ contingency table. Thus the inertia of the contingency table is equal to Pearson's chi-square of the table adjusting for the sample size.

If only the first two eigenvectors are used to
approximate the value of $d_{ij}$, then $f_{ij}$ can be estimated as in the following:

$$f_{ij} = NP_{i+}p_{+j} + \sqrt{NP_{i+}p_{+j}} (d_{ij})$$

$$= NP_{i+}p_{+j} + N/NP_{i+}p_{+j} (\sqrt{\lambda_1 u_{i1}v_{j1}} + \sqrt{\lambda_2 u_{i2}v_{j2}})$$

The above formula is also known as the reconstitution formula. Now each of the eigenvalues is associated with a principal axis. The common practice is to take the first two principal axes to form a plane onto which the rows and columns of matrix $D$ can be projected. The coordinate of row $i$ on the $k$th principal axis is given by

$$\text{coord} (i, k) = \sqrt{\frac{\lambda_k}{\Pi_i}} v_{ik},$$

while that of column $j$ on the same principal axis is given by

$$\text{coord} (j, k) = \sqrt{\frac{\lambda_k}{\Pi_j}} u_{jk}.$$  

These row and column coordinates are related to each other by the following transition equations:

$$\text{coord} (i, k) = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{c} \frac{p_{ij}}{\Pi_j} \text{coord} (j, k)$$

and

$$\text{coord} (j, k) = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{r} \frac{p_{ij}}{\Pi_i} \text{coord} (i, k),$$

satisfying the conditions that
\[ \sum_i p_{i} \, \text{coord}(i,k) = 0 \]

and

\[ \sum_j p_{j} \, \text{coord}(j,k) = 0. \]

By mapping the row and column categories onto the plane formed by, say, the first two principal axes using the coordinate formula given above, a graphical representation of the two-way contingency table is resulted. It provides a portrayal of the dependence of two categorical variables in a visual way. Because of the relationship that links the rows to the columns in the transition equations on each of the principal axis, a row category will have a coordinate that is numerically close to those column categories for which the proportions, \( p_{ij}/p_{i*} \), are large. This fact justifies the representation of both the row and column categories on the same principal axes. However, in this setting, the distance between a row and a column category is not meaningful at all, since it is the collection of all column coordinates that define the coordinate of a row category by virtue of the transition equation and vice versa. Thus proximity between a column and a row has to be interpreted with extreme care. For example, Everitt (1988) suggests the use of such phrases as "the occurrence of the
row category with respect to the column category is proportionally higher than with respect to all the other column categories" as a way to depict proximity between categories across variables. For a complete discussion of correspondence analysis as a whole, please refer to Greenacre (1984).
APPENDIX C

CORRESPONDENCE PLOTS FOR THE STUDY OF
FRESHMEN'S CONSUMPTION OF ALCOHOL

132
Correspondence plot of the frequency of drink and the Greek variables

- drink
- Greek
- none
- not Greek
- 4+ 3 2

Axes:
- Axis 1 (28.5%)
- Axis 2 (20.0%)
Correspondence plot of the frequency of drink and the gender variables.
Correspondence plot of the frequency of drink and the g.p.a. variables

axis 1 (11.0%)
axis 2 (8.98%)

- drink
- g.p.a.
Correspondence plot of the frequency of drink and the age variables.
Correspondence plot of the frequency of drink and the Greek by gender variables

- drink
- Greek_gender
Correspondence plot of the frequency of drink and the Greek by g.p.a. variables

- drink
- Greek g.p.a.
Correspondence plot of the frequency of drink and the Greek by age variables

- G G < 10
- never n G
- < 10 n G
- 1 6 - 17
- m
- n G 12 - 13
- / n G 12 - 13
- B G 18
- none
- \$
- •
- 0
- 1
- 2
- 3
- 4
- 5
- 6

axis 2 (7.41%)

axis 1 (8.78%)

- drink
- Greek age
APPENDIX D
CORRESPONDENCE PLOTS FOR THE STUDY OF
PRODUCTIVITY OF TEACHER EDUCATORS
Correspondence plot of the presentation and the master's thesis variables.
Correspondence plot of the presentation and the g.p.a. variables
Correspondence plot of the presentation and the bachelor's major variables.
Correspondence plot of the presentation and the master's major variables

axis 1 (22.1%)

axis 2 (20.0%)

- presentation
- Education major

not Ed.
none
2
3
4+
Correspondence plot of the presentation and the age at Ph.D. variables.

- Presentation
- Ph.D. Age

- Axis 1 (16.1%)
- Axis 2 (15.6%)
Correspondence plot of the presentation and the full time enrollment variable.
Correspondence plot of the presentation and the teaching experience variables

- Presentation
- Teaching experience
Correspondence plot of the presentation and the decision to become teacher educator variables


Pearson, K. (1900). On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos. Mag.*, Series 5, 50, 157-175.


