DETERMINATION OF STUDENTS' ACADEMIC PERFORMANCE USING CANONICAL CORRELATION AND FACTOR ANALYSES

BY

ABDULSALAM, LUKMANB.Sc. (UDUS 2006)
M.SC/SCIEN/06819/08-09

A THESIS SUBMITTED TO THE POSTGRADUATE SCHOOL, AHMADU BELLO UNIVERSITY, ZARIA, NIGERIA

IN PARTIAL FULFILMENT FOR THE AWARD OF MASTERS OF PHILOSOPHY IN STATISTICS

DEPARTMENT OF MATHEMATICS
AHMADU BELLO UNIVERSITY
ZARIA

MARCH, 2014

DECLARATION

I, Lukman AbdulSalam hereby declare that this research thesis was carried out by me. It has never been presented for any academic degree or diploma in any institution of learning.

Lukman AbdulSalam Date

CERTIFICATION

This thesis titled; "Determination of Students' Academic Performance Using Canonical Correlation And Factor Analyses" meets the regulations governing the award of the degree of Master of Science in Statistics of Ahmadu Bello University, and is approved for its contribution to knowledge and literary presentation.

Prof. S.U. Gulumbe Major Supervisor	Date
Dr. H.G. Dikko Minor Supervisor	Date
Dr. A. D. Koko External Supervisor	Date
Dr. B. Sani Head of Department	Date
Prof. A.A. Joshua Dean, School of Postgraduate Studies	Date

ACKNOWLEDGEMENT

Most importantly, I owe all what I have been obtaining to Allah, Who has always been shining my life. Now I can only reaffirm the following words – “Allah, whatever grace we have, it is from You. No deity worthy of being worshiped except You”.

I would like to thank my supervisors, Prof. S. U. Gulumbe for his guidance and the insight he provided throughout this research thesis and Dr. H.G. Dikko for his support and encouragement on this research thesis and also for reading the whole manuscript and making all necessary corrections. His dedication to research and patience for students will always be example for me to follow. My gratitude also go to all staff of the Department of Mathematics, most importantly Mr. Chibuke Nnamani for his valuable contribution towards making this work a success.

Thanks to my parents who have been extremely understanding and supportive of my studies. I feel very lucky to have parents that share my enthusiasm for academic pursuit, and my brothers and sisters, who are my best friends. I owe myprecious wife;Malama Royhana Akibu Muhammad and daughters, Rumaysa and Nusayba, special thanks for their unquantifiable patience and encouragement over the years.

ABSTRACT

In this study, we tried to evaluatestudents' academic performance through the use of the relationship between mathematical and less-mathematical subjects and also tried to verify the subjects that contribute significantly to the variation among them. Canonical correlation analysis was employed to analyse the relationship between mathematical and less-mathematical subjects and to also test the significance of canonical variate and the homogeneity of variance among the variables obtained with the use of Wilk's Lambda and Bartlett's test respectively. Factor analysis was used to investigate the variability among the subjects and find out the variables that contribute significantly to the percentage of variance obtained. The data used is the Senior Secondary Certificate Examination results of Command Secondary School, Kaduna, conducted by the National Examination Council in the year 2008, 2009 and 2010 respectively. The data consists of results of 90 students in eight subjects out of nine registered subjects by each student. Two sets were formed; set-1 which consists of Mathematics, Chemistry and Physics; was classified as the mathematical subjects, while set-2 which consists of Economics, English Language, Biology, Geography and Hausa Language; was classified as less-mathematical subjects. The data wasanalyzed and structured into mathematical and less-mathematical data using the NCSS 2007 package. The purpose of the structuring was that of gathering information needed to quantify elements which were considered in order to construct determinant factors and to as well check which of the variables contributed significantly to the variance.Our results showed that less-mathematical subjects have significant impact on determining students' academic performance.Three canonical roots were obtained and two are statistically significant showing a strong correlation between the two sets. Four factors were considered and it shows that less-mathematical subjects contribute significantly to the

variation among the variables.Mathematical subjects are directly related, so also less-mathematical subjects. However, mathematical subjects are inversely related to less-mathematical subjects, which means, an increase in the students' performance in mathematical subjects will lower the performance in less-mathematical subjects.

TABLE OF CONTENT

LIST OF TABLES

CHAPTER ONE

General Introduction

1.1 Introduction

Multivariate Statistics is a useful set of methods for analyzing a large amount of information in an integrated frame, focusing on the simplicity (Simon, 1969) and latent order (Wheatley, 1994) in seemingly complex array of variables. Benefits of using multivariate statistics include: Expanding sense of knowing, allowing rich and realistic research designs and having flexibility built on similar univariate methods.

Canonical correlation analysis is the most generalized member of the family of multivariate statistical techniques. It is directly related to several dependence methods, similar to regression; canonical correlation's goal is to quantify the strength of the relationship, in this case between the two sets of variables. It also resembles discriminant analysis in its ability to determine independent dimension for each variable set, in this situation with the objective of producing the maximum correlation between the dimensions. Thus, canonical correlation identifies the optimum structure or the dimensionality of each variable set that maximizes the relationship between dependent and independent variable sets(Anastasi and Urbina, 1997; Harlow, 2005).

Numerous studies, such as those carried out by Fullana (1995) and Montero (1990)have sought to understand the factors which account for low or high achievements. Studies seeking to identify what determines academic failure or success frequently appear as a reaction to conditions of change, such as plans for educational reform, or in response to critical situations: the National Examination Council (NECO) result, 2010, shows that less than twenty percent of those who sat for the examination made their subject combinations that can guarantee them admission into the tertiary institutions.

In general, the various studies which attempt to explain academic failure or success do so beginning with the three elements that intervene in education: parents (family causal factors), teachers (academic causal factors), and students (personal causal factors). Among personal variables most studied are motivation and self-concept. Motivation is considered to be the element that initiates the subject's own involvement in learning: when a student is strongly motivated, all his effort and personality are directed towards the achievement of a specific goal, thus bringing to bear all his or her resources. Some authors have found that subjects themselves attribute low/high performance to low/highability and to luck (Valle, 1999), and an improvement in performance to motivation (task goal orientation), to self-regulating behaviors, and to competence as a function of task characteristics (Slater, 2002). In recent research, positive correlations were found between the value given to the task and the perceptions of auto-efficacy and performance (Chia, 2002). However, in a theoretical review, Fuente (2002) showed how there has been a branching off toward the study of academic goals, to the detriment of those of a social nature, even though these have been shown to be especially important in the most disadvantaged social contexts.

A frequently applied paradigm in analyzing data from multivariate observations is to model the relevant information (represented in a multivariate variable X) as coming from a limited number of latent factors. Because each factor may affect several variables in common, they are known as "common factors". Each variable is assumed to depend on a linear combination of the common factors, and the coefficients are known as loadings. Each measured variable also includes a component due to independent random variability, known as "specific variance" because it is specific to one variable.

In a survey on household consumption, for example, the consumption levels, X, of p different goods during one month could be observed. The variations and covariations of the p components of X throughout the survey might in fact be explained by two or three main social behavior factors of the household. For instance, a basic desire of comfort or the willingness to achieve a certain social level or other social latent concepts might explain most of the consumption behavior (Hardle and Simar, 2003).

Complex multivariate data structures are better understood by studying low-dimensional projections. For a joint study of two data sets, we may ask what type of low-dimensional projection helps in finding possible joint structures for the two samples (Hardle and Simar, 2003).

Bartholomew *et al* (2008) described Factor Analysis as a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors. For example, it is possible that variations in four observed variables, mainly reflect the variations in two unobserved variables. Factor analysis searches for such joint variations in response to unobserved latent variables. The observed variables are modelled as linear combinations of the potential factors, plus "error" terms. The information gained about the interdependencies between observed variables can be used later to reduce the set of variables in a dataset. Computationally this technique is equivalent to low rank approximation of the matrix of observed variables.

Shen *et al* (2009) stated that, the canonical correlation analysis is a standard tool of multivariate statistical analysis for discovery and quantification of associations between two sets of variables. Specifically, this analysis allows us to investigate the relationship between two sets of variables. For example, an educational researcher may want to compute the simultaneous relationship between three measures of scholastic ability with five measures of successes in school. A medical researcher may want to study the relationship of various risk factors to the development of a group of symptoms.

This study is carried out to look at evidence of influence as to how students' preference and extreme rapt attention to some subjects can affect other subjects with less or no preference and attention. It is believed that in doing this, one may be able to adequately counsel students on how to prepare for examinations.

1.2 Statement of the Problem

Various researches which focus on management had been carried out on Canonical Correlation, such as; "Study of Asset Liability Management in Indian Banks, Canonical Correlation Analysis" by Ranjan and Nallari,(2005). In fact, it is rarely applied on the outcome of students' academic performance with a view to analyzing mathematical subjects to establish the relationship between the mathematical and less-mathematical subjects studied in secondary schools. Also, numerous studies had been carried out using Factor Analysis to focus on health and socio-economic issues while we have little to show for analysis of students' academic performance with a view to analyzing mathematical and less-mathematical subjects studied in secondary schools.

In this thesis, we have tried to use a combination of both Canonical Correlation and Factor Analysis to determine students' academic performance of Command Secondary School, Kaduna using mathematical and less mathematical subjects.

1.3 Background to the Study

Predicting students' performance in an academic program is a difficult but useful undertaking. Many studies on predicting academic performance such as one carried out by Mustapha *et al*(2010) and Wooten (1998) have been carried out,using differentpossible influencing factors. While some looked at the teaching/learning processes, good atmosphere for learning and students' behaviors, others looked at the examination outcomes of the students.

Most of the observable phenomena in the empirical sciences are of a multivariate nature. In financial studies, assets in stock markets are observed simultaneously and their joint development is analyzed to better understand general tendencies and to track indices. In medicine, recorded observations of subjects in different locations are the basis of reliable diagnoses and prescription. In quantitative marketing, consumer preferences are collected in order to construct models of consumer behavior. The underlying theoretical structure of these and many other quantitative studies of applied sciences are multivariate.

Factor analysis has provoked rather turbulent controversies throughout its history. Its modern beginnings lie in the early twentieth century attempts of Karl Pearson, Charles Spearman and others to define and measure "intelligence". Because of this early association with constructs such as intelligence, factor analysis was nurtured and developed primarily by scientists interested in psychometric measurement. The advent of high-speed computers has generated a renewed interest in the theoretical and computational aspects of factor analysis. Most of the original techniques have been abandoned and early controversies resolved in the wake of recent developments. It is still true that each application of the technique must be examined on its own merits to determine its success (Richard and Dean, 1992)

Although the numerical methods required for canonical correlation analysis are much more complex than those required for computing a bivariate coefficient, canonical correlation analysis can be conceptually understood in terms familiar to bivariate analysis. Canonical correlation analysis computes to derive variables U and V, such that the correlation between U and V is as large as possible. Canonical correlation analysis computes several such relations between independent and dependent variables. Each relation indicates a distinct pattern that exists in the data. Canonical correlation analysis reduces each of these patterns to derive variables – the canonical *U* and *V* variables. This means, for example that each such relation can be visually inspected using a familiar bivariate scatter diagram. The largest canonical correlation corresponds to the strongest relation between independent and dependent variables. Subsequent canonical correlations correspond to relations of decreasing strength. The significance of this feature for human factor research is that we often find different response patterns under different environmental conditions.

1.4 Aim and Objectives

The aim of this research work is to evaluate Students' Academic Performance through the use of the relationship between mathematical and less-mathematical subjects. This aim will be achieved through the following objectives:

i. To determine the relationship between mathematical and less-mathematical subjects, using Canonical Correlation.

ii. To investigate the homogeneity of variance among the subjects using Bartlett's Test.

iii. To find out the variable that contributes significantly to the percentage of variance using Factor Analysis.

1.5 Scope and Limitation of the Study

The study is centered at analyzing some SSCE results from Command (Day) Secondary School, Kaduna conducted by the National Examination Council (NECO) in the years 2008, 2009 and 2010 respectively. Linear Canonical Correlation and Factor

Analyseswere used for the study. Bartlett's test is used to test for the homogeneity of variance of the variables.

The data used (result of the students) were not individually verified to ascertain if the scripts marked by the examination council are error free. It is also a secondary data and cannot be established if there are different factors that influence the final result of each student either during the course of writing the examination or while marking the scripts by the assessors. The data is also from a single school due to the refusal of many secondary schools to release their students' Senior Secondary Certificate Examination (S.S.C.E) results, for fear of it being exposed to the general public.

1.6 Definition of Terms and Concepts Used

1.6.1 Matrix

A rectangular array of numbers enclosed by a pair of brackets such as

$A = \begin{bmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ z_1 & z_2 & z_3 \end{bmatrix}$ and $B = \begin{bmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{bmatrix}$ each is called a matrix and both are called matrices. The matrix $[x_1 \quad x_2 \quad x_3]$ is called a row vector and $[x_1 \quad x_2 \quad x_3]'$ is called a column vector.

1.6.2 Eigenvalue and Eigenvector

Let A be a symmetric matrix and if there exists a scalar λ such that $|A - \lambda I| = 0$, then the roots of the determinant equation are called the eigenvalues (or characteristic values or latent roots) of the matrix A. If also there is a vector X such that $(A - \lambda I) = 0$, such that

$X \neq 0$, then the columns of X which satisfy the equation are called the eigenvectors (or characteristic root) of A.

1.6.3 Correlation

A measure of the degree of a relationship between two sets of variables is called correlation. The value of the association is called the correlation coefficient, which lies between $-1$and $+1$.

1.6.4 Linear Compound

Suppose X is an n-dimensional random variable with $E(X) = \mu$and $Var(X) = \Sigma$. Then if a is a $(p \times 1)$ vector of constants, $u = a'x$ is a random variable called a linear compound.

1.6.5 Canonical Variates and Canonical Correlation

As in (1.5.10) if there are two linear compounds $u = a'X$ and $v = b'Y$ where X and Y are two sets of variables. Then the values of u and v are called canonical variates and their correlation is called canonical correlation.

1.6.6 Canonical Weight

Canonical weight, also called the canonical function coefficient or canonical coefficient is the coefficient of the linear combinations of X and Y of the linear compounds in (1.5.10). The standardized canonical weights are used to assess the relative importance of individual variables contribution to a given canonical correlation. Canonical coefficients are the standardized weights in the linear equation of variables, which creates the canonical variables.

1.6.7 Canonical Scores

They are values on a canonical variable for a given case, based on the canonical coefficients for that variable. Canonical coefficients multiplied by the standardized scores of the cases and summed to yield the canonical score for each case in the analysis.

1.6.8 Structure Correlation Coefficients (Canonical Loading)

This can also be called canonical factor loading. A structure correlation is the correlation of a canonical variable with the standardized scores of an original input variable. The table of structure correlation is sometimes called the factor structure. The squared structure correlation indicates the contribution made by a given variable to the explanatory power of the canonical variate based on the set of variables to which it belongs.

1.6.9 Canonical Communality Coefficient

This is the sum of the squared structure coefficients for a given variable. The canonical communality coefficient measures how much of a given original variable's variance is reproducible from the canonical variables. As in factor analysis, the original variables with low canonical communality are those for which the model is not working and the researcher may use this information to drop such variables from the analysis.

The canonical variate adequacy coefficient is the average of the entire squared structure coefficient for one set of variables (the dependent or independent set) with respect to a given canonical variable. It is also the percent of variance in the set of variables (e.g. the independent set) extracted by the canonical variate for that set.

1.6.10 Factor Model

Let us assume that $Y_{(p\times1)}$ be the set of observed variables which are linearly related with $f_{(m\times1)}$ a set of unobservable latent variables such that $m < p$. The general model is $Y = \mu + Lf + \varepsilon$. We refer to $f_1, f_2, \dots, f_m$ as common factors and $u_1, u_2, \dots, u_p$ as the specific factors. The quantities are called factor loadings. They are correlations of items (rows) with factors (columns)

1.6.11 Hypotheses

It is often in practice to make decision about a population based on sample information (data). Such decisions are called statistical decisions. To reach a decision, it is useful to make assumptions (or guess) about the population involved. Such assumptions, which may or may not be true, are called statistical hypotheses. They are generally statements about the probability distribution of the population. A null hypothesis is a statement made for the main purpose of rejecting it. If the null hypothesis is rejected then there must be an alternative. This alternative is called alternative hypothesis that will be accepted if the null hypothesis is rejected.

1.6.12 Latent Variable Analysis

Latent variable analysis focuses on three or more variables at least one of which is unobserved. As such, latent variable analysis is appropriate any time a researcher wishes to evaluate a theoretically posited relationship among concepts by the examination of measures designed to measure those concepts such as definition means that virtually all research is "latent variable analysis refers to factor analysis".

CHAPTER TWO

Literature Review

2.1 The Concept of Academic Performance

The very concept of academic failure varies in its definition. Bonaci*et al.* (2010) considered academic failure as the situation in which, student does not attain the expected achievement according to his or her abilities, resulting in an altered personality which affects all other aspects of life. Similarly, Tapia (2002) noted that, while the current Educational System perceives that the student fails if he or she does not pass, more appropriate for determining academic failure is whether the student performs below his or her potential.

Wooten (1998) undertook a study of 271 students taking introductory accounting at a major southeasternAmerican university of which there were 74 students equal to or older than25 years age identified as non-traditional, while 127 students were under 25 years age identified as traditional. He found that, for thetraditional cohort grade history, motivations and family responsibilities all influenced the amount ofeffort these students made. However, neither extracurricular activities nor work responsibilities influenced their effort. However for the non-traditional students, motivation was the only variablethat significantly influenced effort. Neither grade history nor extracurricular activities, nor workresponsibilities, nor family responsibilities had an effect on motivations. Family activities had asignificant negative impact on effort for the traditional students, but not for the nontraditional students. It is conjectured by the authors of this paper that these differences in ages may also capture different socio-economic circumstances.

Much research has been done on common predictive factors of academic performance in accountingcourses, including gender, prior knowledge of accounting, academic aptitude, mathematicalbackground, previous working experience, age, class size, lecturer attributes and student effort, asdocumented by Naser and Peel (1998) and Koh and Koh (1999). The findings are not definitive.

McKenzie and Schweitzer (2001) investigated academic, psychosocial, cognitive and demographic predictors of academic performance to improve interventions and support services for student at risk of academic problems. They recommended implementing stringent record keeping procedures at the university level to enable researchers to fully examine the relationship between age, previous academic performance and university achievement.

Nonis and Hudson (2006) noted that the Higher Education Research Institute at UCLA's GraduateSchool of Education has found that since 1987 the time students spend studying outside of class has declined each year, with only 47% spending six or more hours per week studying outside of classcompared with 34% in 2003. This corresponds with findings of Gose(1998) who found an increase in the number of students employed with 39% of students working 16 or more hours per week in 1998 compared with 35% working in 1993. Nonis and Hudson (2006) identified a need for empirical research to determine the impact of student work on academic performance, and its impact on the design of academic programs. Their study found a lack of evidence for a direct relationship between time spent working and academic performance.

Bonaci*et al.* (2010) on their part, tested the influence of three established dimensions on students' academic performance - self acknowledged academic performance namely; students, teaching / learning environment and professors. Considering an individual perspective for each of the three dimensions, they find that students' performance is only influenced by their own dimension, while the academic environment and professor do not influence their assumed / acknowledged performance.

2.2 Canonical Correlation Analysis of Students' Performance

Canonical Correlation Analysis is a type of multi-variate statistical analysis, first described by Hotelling (1936). It is currently being used in fields like chemistry, biology, meteorology, demography, artificial intelligence, cognitive science, political science, sociology, psychometrics, educational research, economics and management science to analyze multi-dimensional relations between multiple independent and multiple dependent variables.

Canonical correlation analysis is most closely related to multiple regressions and to Principal component analysis. It differs from Principal component analysis in that Principal component analysis makes no distinction between independent and dependent variables. Canonical correlation analysis differs from multiple regressions since it has a multiple dependent variable.

Shafto (1997) said, although Canonical correlation analysis is found in standard textbooks and is available in many statistical computing packages,there are some technical and interpretive problems which prevent its routine use by human factor practitioners which are computation, interpretation and statistical significance. They illustrate the problems and their solutions based on their experience in using canonical correlation in the analysis of a field study of crew-automation interaction in commercial aviation.

Borga (2001) stated in his piece that an important property of canonical correlations is that they are invariant with respect to affine-transformation of the variables. He went further and stated that the number of non-zero eigenvalues are limited to the smallest dimensionality of X and Y (i.e. if $X_{(p\times1)}$and $Y_{(q\times1)}$) such that $p < q$, then maximum number of correlation is p.

Liu (2004) applied canonical correlation analysis to present a method for predictive modeling of anatomic structures which by using this technique, certain anatomical structures, such as tumor distorted structures can be estimated from others by exploring the correlation between them, which has been determined from a set of training samples.

Liu (2004)performed two experiments on cortical surfaces and corpus callosum boundaries in order to evaluate the canonical correlation analysis based estimation method. In the first experiment, the inner cortical surfaces of 60 normal individuals were registered. As the goal of the experiment was to evaluate the estimation method, they masked out part of the anatomy lying within a spherical region and treated the masked anatomy as the unknown anatomy to be estimated. Then the masked anatomy was used as the ground truth to evaluate the estimation method. The results were compared using canonical correlation analysis method with those of three other methods; the principal component regression, multiple linear regressions, a principal component analysis-head method. The four methods were tested using different numbers of training samples variously sized anatomic masked, and various positions for those masks. Of the four methods, canonical correlation analysis produced the best results with a sample of 100.

Ranjan and Nallari (2005) used canonical correlation analysis to analyze asset liability management (ALM) as a key driver of a bank's profitability and long-term sustainability. It required assessment of various types of risks and altering the asset liability portfolio to manage risk. Their objective was to study the portfolio Matching Behavior of Indian Banks in terms of nature and strengths of relationship between Assets and Liability and analyzed the impact of ownership over Asset Liability management in Banks. They studied the nature and strength of relationship between the assets and liabilities. The study showed that most of the banks had a good Asset Liability Management (ALM) framework in place. There was a strong relationship between the Fixed Assets and Network for all the banks. They also discovered that the banks (based on ownership) differ on some of the other aspects like their risk appetite.

2.3 Factor Analysis of Students' Performance

Factor analysis is a form of exploratory multivariate analysis that is used to either reduce the number of variables in a model or to detect relationships among variables.

Richard and Dean (1992) declared that the essential purpose of factor analysis is to describe, if possible, the covariance relationships among many variables in terms of a few underlying but unobservable random quantities called factors. Basically, the factor model is motivated by the following argument; suppose variables can be grouped by their correlations. That is, all variables within a particular group are highly correlated among themselves but have relatively small correlations with variables in a different group. It is conceivable that each group of variables represents a single underlying construct or factor that is responsible for the observed correlations. For example, correlations from the group of test scores in classes. French, English, Mathematics and Music collected by Spearman suggested an underlying "intelligence" factor. A second group of variables representing physical-fitness scores, if available, might correspond to another factor. It is this type of structure that factor analysis seeks to confirm.

Ballester *et al* (2004) compared the performance of principal component analysis and the principal factor analysis. According to them, principal component analysis is recommended for dimensional reduction, whereas principal factor analysis is adapted for the study of structures in the data. In particular, Widaman states, "principal component analysis should not be used if a researcher wishes to obtain parameters reflecting latent structures or factors".

Golding and Donaldson (2006) stated that the use of performance in first year computer science course is a possible factor, which may determine academic performance. They also showed that gender and age have no significant correlation as predictive factors. Guney (2009) also discusses the idea of examination structure and emphasizes that students, care about the way the examination questions are asked.

Mustata*et al.* (2010) developed a review on influential factors of accounting students' academic performance concluding that an overall consensus with regard to determinants of academic success in accounting wasn't yet achieved, rather inconclusive results having been produced.

2.4 Other Statistical Analysis of Students' Performance

Jordan and Suzanne (2006) conducted a study to test empirically the hypotheses found in the literature regarding impact of moral orientation on the counseling process. The correlation between moral orientation parables (match, mismatch, care and justice) and counselor effectiveness ratings were examined. Naturalistic study was reduced using video footage from real counseling training session. The results suggested that therapist effectiveness was not perceived differentially by clients according to whether or not the therapists matched the client's voice. Clients also found the therapists effective whether the therapists spark predominately in the care or justice voice.

Chamillard (2006) used statistical analysis techniques to predict student performance in a particular course. Observations from the analysis provide useful insight into the relationships between courses.

Superby*et al.* (2006) and Vandamme*et al.* (2007) studied correlations of various parameters such as attendance, estimated chance of success, previous academic experience and study skills. They found out that changing process factors during a student's stay at the university plays a large part in academic performance. In addition they experimented on predicting students' performance using decision tree, neural networks and linear discriminant analysis. The rates of prediction obtained were not particularly good due to the difficulty to classify students into 3 groups, namely, high risk, medium risk and low risk, before the first university examinations.

CHAPTER THREE

Materials and Methodology

3.1 Data Collection

The data used for this analysis was collected from the outcome of grades scored by students of Command Day Secondary School, who wrote the 2010 senior secondary school certificate examination conducted by the National Examination Council (NECO). See appendix for detail.

3.2 Mathematical computation of Canonical Correlation Analysis

Anderson (1958) gave a detailed mathematical concept of canonical correlation analysis. Let $\boldsymbol{X}$ be a q-dimensional random vector and $\boldsymbol{Y}$be a p-dimensional random vector. Suppose that $\boldsymbol{X}$and $\boldsymbol{Y}$have mean $\boldsymbol{\mu}$and $\boldsymbol{v}$respectively and that

$$\boldsymbol{E}[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})'] = \sum\nolimits_{\mathbf{11}} \qquad \mathbf{(3.1)}$$

$$\boldsymbol{E}[(\boldsymbol{y}-\boldsymbol{v})(\boldsymbol{y}-\boldsymbol{v})'] = \sum\nolimits_{\mathbf{22}} \qquad \mathbf{(3.2)}$$

Let us now consider the two linear combinations

$$\boldsymbol{g} = \boldsymbol{a}'\boldsymbol{x} \qquad \mathbf{(3.3)}$$

and

$$\boldsymbol{f} = \boldsymbol{b}'\boldsymbol{y} \qquad \mathbf{(3.4)}$$

The correlation between $\boldsymbol{g}$ and $\boldsymbol{f}$ is defined as shown below

$$\boldsymbol{\rho}\,(\boldsymbol{a},\boldsymbol{b}) = \frac{\boldsymbol{a}'\sum_{\mathbf{12}}\mathbf{b}}{[(\boldsymbol{a}'\sum_{\mathbf{11}}\boldsymbol{a})(\mathbf{b}'\sum_{\mathbf{22}}\mathbf{b})]^{\mathbf{1/2}}} \qquad (\mathbf{3.5})$$

Our canonical variables are

$$\boldsymbol{g} = \boldsymbol{a}_0' \boldsymbol{\Sigma}_{11}^{-1/2} \boldsymbol{X} \qquad \textbf{(3.6)}$$

and

$$\boldsymbol{f} = \boldsymbol{b}_0' \boldsymbol{\Sigma}_{22}^{-1/2} \boldsymbol{Y} \qquad \textbf{(3.7)}$$

which can be used to find the canonical correlation coefficient ($\boldsymbol{corr}(\boldsymbol{g}, \boldsymbol{f})$), which is the measure of the association between $\boldsymbol{g}$ and $\boldsymbol{f}$**.**

3.3 Mathematical Computation of Factor Analysis

Multiple factor is based on the premise that a large number of questionnaire items could be reduced to only a few dimensions. The multiple factor approach emphasized the goal of extracting a maximum amount of variance from a correlation.

Consider the general factor model

$$\boldsymbol{X}_{p\times 1} = \boldsymbol{\mu}_{p\times 1} + \boldsymbol{L}_{p\times m} \boldsymbol{f}_{p\times 1} + \boldsymbol{\varepsilon}_{p\times 1} \qquad \textbf{(3.8)}$$

The regression coefficient $\boldsymbol{l}$(the partial slopes) for all of these multiple regressions are called i^{th}variable and the j^{th}factor, which could be collected in a matrix as shown below

$$\boldsymbol{L} = \begin{pmatrix} \boldsymbol{l}_{11} & \cdots & \boldsymbol{l}_{1m} \\ \vdots & \ddots & \vdots \\ \boldsymbol{l}_{p1} & \cdots & \boldsymbol{l}_{pm} \end{pmatrix} \qquad \textbf{(3.9)}$$

3.3.1 Orthogonal Factor Model

The aim of factor analysis is to explain the outcome of $\boldsymbol{p}$ variables in the data matrix χ using fewer variables, *the so-called factors*. Ideally all the information in χ can be reproduced by a smaller number of factors. These factors are interpreted as latent (unobserved) common characteristics of the observed $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{\boldsymbol{p}}$. The case just described occurs whenever y observed $\boldsymbol{\mathcal{X}} = (\boldsymbol{\mathcal{X}_1}, \dots, \boldsymbol{\mathcal{X}_p})^{\top}$ can be written as

$$\boldsymbol{\mathcal{X}_j} = \sum_{\ell=1}^{k} \boldsymbol{qj\ell f\ell} + \boldsymbol{\mu_j}, \boldsymbol{j} = \boldsymbol{1}, \dots, \boldsymbol{p}. \qquad \textbf{(3.10)}$$

Here $\boldsymbol{f\ell}$, for $\boldsymbol{\ell} = \boldsymbol{1}, \dots, \boldsymbol{k}$denotes the factors. The number of factors, $\boldsymbol{k}$**,** should always be much smaller than $\boldsymbol{p}$**.**

We can create a representation of the observations that is similar to the one in equation (3.11) by means of principal components, but only if the last $\boldsymbol{p-k}$ eigenvalues corresponding to the covariance matrix are equal to zero. Consider a p-dimensional random vector X with mean μ and covariance matrix $\text{Var}(X) = \sum$. A model similar to equation (3.10) can be written for X in matrix notation,

$$\boldsymbol{X = QF + \mu} \qquad \textbf{(3.11)}$$

where$F$ is the $\boldsymbol{k}$ $-$dimensional vector of the $\boldsymbol{k}$factors.

The spectral decomposition of $\sum$ is given by $\Gamma\Lambda\Gamma^{\top}$. Suppose that only the first $\boldsymbol{k}$eigenvalues are positive, i.e., $\lambda_{k+1} = \cdots = \lambda_p = 0$. then the (singular) covariance matrix can be written as

$$\sum = \sum_{\boldsymbol{\ell=1}}^{\boldsymbol{k}} \boldsymbol{\lambda_\ell \gamma_\ell \gamma_\ell}^{\top} = (\boldsymbol{\Gamma_1 \Gamma_2}) \begin{pmatrix} \boldsymbol{\Lambda_1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Gamma_1}^{\top} \\ \boldsymbol{\Gamma_2}^{\top} \end{pmatrix}. \qquad (\boldsymbol{3.12})$$

Note that the covariance matrix of equation (3.12) can be written as

$$\sum = \boldsymbol{E(X-\mu)(X-\mu)}^{\top} = \boldsymbol{QE(FF^{\top})Q^{\top}} = \boldsymbol{QQ}^{\top} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{k}} \boldsymbol{\lambda_j \gamma_j \gamma_j^{\top}} \qquad (\boldsymbol{3.13})$$

It is common praxis in factor analysis to split the influences of the factors into common and specific ones. There are, for example, highly informative factors that are common to all of the components of $\boldsymbol{X}$and factors that are specific to certain components. The factor analysis model used in praxis is a generalization of equation (3.12):

$$\boldsymbol{X} = \boldsymbol{QF} + \boldsymbol{U} + \boldsymbol{\mu}, \qquad \textbf{(3.14)}$$

where$Q$is a $(p \times k)$matrix of the (non-random) loading of the common factors $F(k \times 1)$ and U is a $(p \times 1)$ matrix of the (random) specific factors. It is assumed that the factor variables F are uncorrelated random vectors and that the specific factors are uncorrelated and have zero covariance with the common factors. More precisely, it is assumed that:

$$\boldsymbol{EF} = \boldsymbol{0},$$

$$\boldsymbol{Var}(\boldsymbol{F}) = \boldsymbol{\mathcal{I}_k},$$

$$\boldsymbol{EU} = \boldsymbol{0}, \qquad \textbf{(3.15)}$$

$$\boldsymbol{Cov}(\boldsymbol{U_i}, \boldsymbol{U_j}) = \boldsymbol{0}, \quad \boldsymbol{i \neq j}$$

$$\boldsymbol{Cov}(\boldsymbol{F}, \boldsymbol{U}) = \boldsymbol{0}.$$

The generalized factor model (3.14) together with the assumptions given in (3.15) constitutes the orthogonal factor model.

Note that equation (3.14) implies for the components of $X = (X_1, \ldots, X_p)^\intercal$ that

$$\boldsymbol{X_j} = \sum_{\boldsymbol{\ell=1}}^{\boldsymbol{k}} \boldsymbol{q_{j\ell} F_\ell} + \boldsymbol{U_j} + \boldsymbol{\mu_j}, \qquad \boldsymbol{j = 1, \ldots, p}. \qquad (\mathbf{3.16})$$

Using equation (3.15) we obtain $\sigma_{X_j X_j} = Var(X_j) = \sum_{\ell=1}^{k} q_{j\ell}^2 + \psi_{jj}$.The quantity $h_j^2 = \sum_{\ell=1}^{k} q_{j\ell}^2$ is called the communality and ψ_{jj} the specific variance. Thus the covariance of X can be rewritten as

$$\sum = \boldsymbol{E}(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^{\top} = \boldsymbol{E}(\boldsymbol{QF}+\boldsymbol{U})(\boldsymbol{QF}+\boldsymbol{U})^{\top}$$

$$= \boldsymbol{QE}(\boldsymbol{FF}^{\top})\boldsymbol{Q}^{\top} + \boldsymbol{E}(\boldsymbol{UU}^{\top}) = \boldsymbol{Q}\,\boldsymbol{Var}(\boldsymbol{F})\boldsymbol{Q}^{\top} + \boldsymbol{Var}(\boldsymbol{U})$$

$$= \boldsymbol{QQ}^{\top} + \boldsymbol{\psi}\,. \qquad (\mathbf{3.17})$$

3.3.2 Estimating Factor Scores

Factor Scores(also called component scores in Principal Component Analysis), are the scores are the scores of each case (row) on each factor (column). To compute the factor score for a given case of a given factor, one takes the case's standardized score on each variable, multiplies by the corresponding factor loading of the variable for the given factor, and sums these products. Therefore, given the factor model

$$\boldsymbol{Y_i} = \boldsymbol{\mu} + \boldsymbol{Lf_i} + \boldsymbol{\varepsilon_i}, \ \ \forall \boldsymbol{i} \qquad \textbf{(3.18)}$$

there are many methods for computing factor scores $f = (f_1, f_2, \dots, f_m)$. For this study, we used the method of Ordinary Least Squares. We have

$$\sum_{i=1}^{p} \boldsymbol{\varepsilon_{ij}^2} = \sum_{i=1}^{p} (\boldsymbol{y_i} - \boldsymbol{\mu} - \boldsymbol{l\,f_i})^2 = (\boldsymbol{y_i} - \boldsymbol{\mu} - \boldsymbol{lf_i})(\boldsymbol{y_i} - \boldsymbol{\mu} - \boldsymbol{lf_i})' \qquad (\mathbf{3.19})$$

This takes the form of a least square regressions, only that we have known our L, but wish to estimate the

$$\hat{\boldsymbol{f}}_i = (\boldsymbol{LL})^{-1}\hat{\boldsymbol{L}}(\boldsymbol{y}_i - \boldsymbol{\mu}) \qquad \textbf{(3.20)}$$

Estimated factor loadings are:

$$\hat{\boldsymbol{f}}_i = (\boldsymbol{LL})^{-1}\hat{\boldsymbol{L}}(\boldsymbol{y}_i - \boldsymbol{\mu}) \qquad \textbf{(3.21)}$$

Using the Principal Component method with the un-rotated factor loadings, this yields

$$\hat{\boldsymbol{f}}_i = \begin{pmatrix} \frac{\mathbf{1}}{\sqrt{\hat{\boldsymbol{\lambda}}_1}}\hat{\boldsymbol{e}}_1((\boldsymbol{Y}_1 - \overline{\boldsymbol{y}})) \\ \frac{\mathbf{1}}{\sqrt{\hat{\boldsymbol{\lambda}}_2}}\hat{\boldsymbol{e}}_2((\boldsymbol{Y}_2 - \overline{\boldsymbol{y}})) \\ \vdots \\ \frac{\mathbf{1}}{\sqrt{\hat{\boldsymbol{\lambda}}_m}}\hat{\boldsymbol{e}}_m((\boldsymbol{Y}_m - \overline{\boldsymbol{y}})) \end{pmatrix} \qquad (\mathbf{3.22})$$

3.4 Method of Computation of Canonical Coefficient

Consider two sets of variables $Y = Y_{n\times m}$and $X = X_{n\times q}$,where $m \leq q$.

Construct the linear combinations i.e.

$$\boldsymbol{g} = \boldsymbol{a}'\boldsymbol{X} \qquad \textbf{(3.23)}$$

and

$$\boldsymbol{f} = \boldsymbol{b}'\boldsymbol{Y} \qquad \textbf{(3.24)}$$

such that Γ_{gf} is a maximum.

The XY matrix is

$$\mathbf{XY} = \begin{bmatrix} \boldsymbol{X_{11}} \; \cdots & \boldsymbol{X_{1q}Y_{11}} \cdots & \boldsymbol{Y_{1p}} \\ \boldsymbol{X_{11}} \cdots & \boldsymbol{X_{2q}Y_{21}} \cdots & \boldsymbol{X_{2p}} \\ \vdots & \vdots\vdots & \vdots \\ \boldsymbol{X_{n1}} \cdots & \boldsymbol{X_{mp}Y_{n1}} \cdots & \boldsymbol{X_{np}} \end{bmatrix} \qquad \textbf{(3.25)}$$

The correlation matrix of the above data and its partition is as shown in equation (3.26) below.

Let R denote the correlation matrix, then the partitioned correlation matrix R, is given by

$$\mathbf{R} = \left(\begin{array}{c|c} \boldsymbol{R_{xx}} & \boldsymbol{R_{xy}} \\ \hline \boldsymbol{R_{yx}} & \boldsymbol{R_{yy}} \end{array} \right) \qquad \textbf{(3.26)}$$

where R_{xx} and R_{yy} are the correlation matrices within X and Y respectively and $R_{yx} = R'_{xy}$ which is the correlation matrix between X and Y. To obtain the eigenvalues and their corresponding eigenvectors, it is required to form a symmetric matrix $\mathrm{R}_{11}^{-\frac{1}{2}}\mathrm{R}_{12}\,\mathrm{R}_{22}^{-1}\mathrm{R}_{21}\,\mathrm{R}_{11}^{-\frac{1}{2}}$. The eigenvalues and the corresponding eigenvectors of $\mathrm{R}_{11}^{-\frac{1}{2}}\mathrm{R}_{12}\,\mathrm{R}_{22}^{-1}\mathrm{R}_{21}\,\mathrm{R}_{11}^{-\frac{1}{2}}$ give the canonical correlations and the canonical coefficients of the independent variable X, we use the equation,

$$\boldsymbol{a_1} = \mathbf{R}_{\mathbf{11}}^{-\frac{1}{2}}\boldsymbol{\ell_1} \qquad \textbf{(3.27)}$$

where ℓ_1 is the corresponding eigenvector for the independent variable.

To find the corresponding eigenvectors for Y, use the equation

$$\boldsymbol{h_1} = \frac{1}{\lambda} \mathbf{R}_{yy}^{-\frac{1}{2}} \boldsymbol{\Sigma}_{yx} \boldsymbol{a_1} \qquad \textbf{(3.28)}$$

We can now obtain the canonical coefficients for the dependent variable Y, by applying the equation

$$\boldsymbol{b_1} = \frac{1}{\lambda} \boldsymbol{\Sigma}_{yy}^{-1} \boldsymbol{\Sigma}_{yx} \boldsymbol{a_1} \qquad \textbf{(3.29)}$$

3.5 Significant Test

The significance test of the canonical correlation is straightforward in principle. Simply stated, the different canonical correlations are tested, one by one, beginning with the largest one. Only those roots that are statistically significant are then retained for subsequent interpretation. First, evaluate the significance of all roots combined, then of the roots remaining after removing the first root, the second root and so on. Bartlett (1937) has outlined procedure for testing the significance of canonical correlations. He defined Lambda (Λ)

$$(\boldsymbol{\Lambda}) = \prod_{i=1}^{q} (\mathbf{1} - \boldsymbol{\lambda}_i), \quad \mathbf{q} < p \qquad (\mathbf{3.30})$$

$$\boldsymbol{\chi^2} = -[\boldsymbol{N} - \mathbf{5}(\boldsymbol{p} + \boldsymbol{q} + \mathbf{1})] \ln \boldsymbol{\Lambda} \qquad \textbf{(3.31)}$$

3.5.1 Wilk’s Lambda Test

Wilk's Lambda test is a test statistic used in multivariate analysis of variance (MANOVA) to test whether there are differences between the means of identified groups of subjects on a combination of dependent variables. It is a direct measure of the proportion of variance in the combination of dependent variables that is unaccounted for by the independent variable (the grouping variable or component). Wilk's Lambda statistic can be transformed (mathematically adjusted) to a statistic which has approximately on F distribution. This makes it easier to calculate p – value. Everitt and Dunn (1991).

The Wilk's Lambda test of hypothesis is given as:

H_o : $\sum_{xy} = 0$, i.e. there is no relationship between the canonical variates.

H_1 : $\sum_{xy} \neq 0$, i.e. there is relationship between the canonical variates.

Test statistic:

$$\boldsymbol{\lambda_1} = \frac{|\boldsymbol{R}|}{|\boldsymbol{R}_{\mathrm{yy}}||\boldsymbol{R}_{\mathrm{xx}}|} \qquad (\mathbf{3.32})$$

where:

$\boldsymbol{\lambda_i}$is the symbol for Wilk's Lambda test

$R$is the correlation between x's and y's

$R_{\mathrm{xx}}$is the correlation between x's

$R_{\mathrm{yy}}$is the correlation between y's

Significance Level:

α = 0.05

Decision rule:

Reject H_o if$p < 0.05$ and otherwise accept. Rencher (2002)

3.5.2 Bartlett's Test

Bartlett's Test is used to test the homogeneity of variance in the factors. Bartlett's Test (Snedecor and Cochran, 1983) is used to test if k group have equal variances. Equal variances across groups or samples are called homogeneity of variance. Bartlett's Test is sensitive to departures from normality. That is, if your groups/samples come from non-normal distributions. Then Bartlett's test may be testing for non-normality. The Bartlett's test of hypothesis is given as:

$H_o : \delta_1 = \delta_2 = \ldots = \delta_k$

$H_1 : \delta_i \neq \delta_j$ *for at least one pair(i, j)*

Test Statistic:

$$T = \frac{(N-k)\ln S_p^2 - \sum_{i=1}^{k}(N_i - 1)\ln S_i^2}{1 + \left(\frac{1}{3(k-1)}\right)\left(\left(\sum_{i=1}^{k} 1/(N_i - 1)\right) - 1/(N-k)\right)} \qquad (3.33)$$

In the above, S_i^2 is the variance of the i^{th}group, N is the total sample size, N_i is the sample size of the i^{th} group, k is the number of groups, and S_p^2 is the pooled variance. The pooled variance is weighted average of the group variance and is defined as:

$$S_p^2 = \sum_{i=1}^{k}(N_i - 1)\, S_i^2/(N-k)\,. \qquad (3.34)$$

Significance Level:

$\alpha = 0.05$

Critical Region: The variances are judged to be unequal if, $T > \chi^2_{(\alpha,k-1)}$

Where $\chi^2_{(\alpha,k-1)}$ is the upper critical value of the chi-square distribution with k-1 degree of freedom and a significance level of α.

3.6 Interpretation

Having obtained the significant canonical correlations and their corresponding canonical coefficients for the two sets, we need to go further and interpret the correlation coefficients. The following relations are used for interpretation. The values of canonical coefficient indicate individual correlation between each of the variables in each set with respect to its canonical variate. For example

$$\boldsymbol{U_1 = a_{11}Y_1 + a_{12}Y_2 + \cdots + a_{1i}Y_i + \cdots + a_{1p}Y_p, (i = 1(1)p).} \qquad \textbf{(3.35)}$$

The constant coefficients $\boldsymbol{a_{11}, a_{12}, \dots, a_{1p}}$are partial correlations between $\boldsymbol{U_i}$and the original variables $\boldsymbol{Y_1, Y_2, \dots, Y_p}$. Each of $\boldsymbol{a_{1p}}$is the correlation between $\boldsymbol{U_i}$and the corresponding $\boldsymbol{Y_p}$ **.**

Canonical structure coefficients look at the relationship between each individual variable in a given set and its composite variate (i.e., canonical variable). For example, suppose $= \boldsymbol{a'Y}$ **,** is the canonical variable of the dependent set, $\boldsymbol{Y}$.

Then,

$$\boldsymbol{Corr(Y,U) = Corr(Y,a'Y) = E(Y-\overline{Y})(a'Y-a'\overline{Y})' = E(Y-\overline{Y})(Y-\overline{Y})'a = \mathrm{R_{yy}}a} \qquad \textbf{(3.36)}$$

Similarly,

$$\boldsymbol{Corr}(\boldsymbol{X}, \boldsymbol{b}'\boldsymbol{X}) = \mathbf{R}_{\mathbf{xx}}\boldsymbol{b} \qquad \textbf{(3.37)}$$

The third relationship is the canonical cross loading. This looks at the relationship between the canonical variate and variables of its opposite set. For example

$$\mathbf{Corr}(\boldsymbol{X}, \boldsymbol{a}'\boldsymbol{Y}) = \boldsymbol{E}(\boldsymbol{X} - \overline{\boldsymbol{X}})(\boldsymbol{a}'\boldsymbol{Y} - \boldsymbol{a}'\overline{\boldsymbol{Y}})' = \mathbf{R}_{\mathbf{xy}}\boldsymbol{a} = \mathbf{R}_{\mathbf{xx}}\boldsymbol{b}\boldsymbol{\rho}. \textbf{ Since} \boldsymbol{\Sigma}_{\mathbf{21}}\boldsymbol{a} = \boldsymbol{\rho}\boldsymbol{\Sigma}_{\mathbf{22}}\boldsymbol{b} \textbf{(3.38)}$$

Similarly,

$$\mathbf{Corr}\left(\boldsymbol{Y}, \boldsymbol{b}'\boldsymbol{X}\right) = \mathbf{R}_{\mathbf{yy}}\boldsymbol{a}\boldsymbol{\rho} \textbf{(3.39)}$$

3.7 Statistical Package used for the Study

The statistical package used on the data is NCSS 2007; it was used for the study of canonical correlation and factor analyses.

CHAPTER FOUR

Results and Discussion

4.1 Introduction

This chapter deals with the analysis and discussion of the data considered in the research work. The Analysis of Canonical Correlation and Factor analysis were adopted.

4.2 Results from the Analysis of the Data Used

Table 4.1: Canonical Correlation Coefficient of Set – 1 and Set – 2

Canonical Functions	Canonical Correlation	Eigenvalues	% of Variance Explained
First Pair of Canonical Variance	0.6681	0.4463	68.2
Second Pair of Canonical Variance	0.4152	0.1724	26.4
Third Pair of Canonical Variance	0.1881	0.0354	5.4

See: Appendix 1, Page 50.

Table 4.2: To test that the Canonical Correlations are zero

S/NO	N	P	q	Df	p-value	$\alpha - value$
1	90	5	3	15	0.0000	0.05
2	90	4	2	8	0.0147	0.05
3	90	3	1	3	0.3841	0.05

See: Appendix 1, Page 50.

Table 4.3: Canonical Loading for Set – 1 and Set – 2

Sets	Subjects	r_1	r_2	r_3

Set – 1	Mathematics	-0.6640	0.8257	-0.0858
	Chemistry	-0.4760	-0.5044	-0.7836
	Physics	-0.2566	-0.5624	0.9136
Set – 2	Biology	-0.3046	-0.3856	0.2011
	Hausa Language	-0.1217	0.1247	-0.7266
	Economics	-0.3095	-0.5224	-0.5802
	English Language	-0.5109	0.9406	-0.1720
	Geography	-0.1941	-0.2652	0.8180

See: Appendix 1, Page 51

Table 4.4: Canonical Cross Loading for Set – 1 and Set – 2

Sets	Subjects	r_1	r_2	r_3
Set – 1	Mathematics	-0.5417	0.2375	0.0234
	Chemistry	-0.4242	-0.2348	-0.0991
	Physics	-0.4147	-0.1791	0.1233
Set – 2	Biology	-0.4493	-0.2098	0.0040
	Hausa Language	-0.1677	0.0934	-0.1137
	Economics	-0.4515	-0.2213	-0.0486
	English Language	-0.5463	0.2273	0.0172
	Geography	-0.4737	0.0253	0.0973

See: Appendix 1, Page 52.

Table 4.5: Bartlett's Test

Bartlett's Test, Approx. chi-square	176.35
Degree of Freedom	28
Probability Level	0.000000
α – value	0.05

See: Appendix 2, Page 53.

Table 4.6: Total Variance Explained by each Component

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Eigenvalue	% of Variance	Cumulative %	Eigenvalue	% of Variance	Cumulative %
1	1.2979	28.73	28.73	1.2979	28.73	28.73
2	1.5311	33.89	62.62	1.5311	33.89	62.62
3	0.4047	8.96	71.57	0.4047	8.96	71.57
4	0.5163	11.43	83.00	0.5163	11.43	83.00
5	0.6779	15.01	98.01	0.6779	15.01	98.01
6	0.1242	2.75	100.76			
7	0.0044	0.10	100.85			
8	-0.0386	-0.85	100.00			

See: Appendix 2, Page 54.

Table 4.7: Communalities Extracted by each Variable

Variables	Initial	Extraction
Mathematics	0.6336	0.2176
Physics	0.4856	0.4502
Chemistry	0.4614	0.3803
Economics	0.6675	0.6411
English Language	0.7902	0.5990
Biology	0.5903	0.5542
Geography	0.6454	0.6351
Hausa Language	0.2781	0.2724

See: Appendix 2, Page 53.

Table 4.8: Factor Loadings showing Correlation between Factors and each variable Factor

Variables	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Mathematics	−0.3531	0.1052	0.1775	0.2243	−0.6443
Physics	−0.1818	0.3176	−0.0583	0.5594	−0.1860
Chemistry	−0.1132	0.5542	0.1679	0.1793	0.0440
Economics	−0.1654	0.7648	−0.1557	0.0677	−0.1585
English Language	−0.7438	0.1672	0.1329	−0.0126	−0.4258
Biology	−0.1331	0.6786	0.1540	0.2286	−0.0729
Geography	−0.7224	0.1949	0.1081	0.2522	−0.0934
Hausa Language	−0.0847	0.0246	0.5141	−0.0175	−0.0750

See: Appendix 2, Page 56.

4.3 Discussion on the Result of the Analysis from the Data Used

Table 4.1 shows the Canonical Correlation of the three canonical variates and their corresponding Eigenvalues. The Eigenvalues of the canonical variates can be tested by employing Wilk's Lambda criterion to test for the significant by using Wilk's Lambda test, Rencher (1998).

Since one of the Canonical Correlation tested is significant i.e. at the first pair of Canonical Correlation, where p – value $<\alpha = 0.05$, it implies that the null hypothesis is rejected, which indicates that one out of the three Canonical Correlation is significantly different from zero.

Now, consider the first canonical variatepair $X_1$and Y_1 with Canonical Correlation Coefficient $r_1 = 0.6681$, so that the proportion of variance common to the first pair of canonical variate is $r_1^2 = 0.4463$ showing about 68.20% of the proportion of variance captured by the first canonical variate.

Similarly, $r_2 = 0.4152$ is the Canonical Correlation Coefficient between the second pair of canonical variate, so $r_2^2 = 0.1724$, which indicates about 26.4% of the proportion of variance captured.

Also, $r_3 = 0.1881$ shows the Canonical Correlation Coefficient between the third pair of canonical variate, so $r_3^2 = 0.0354$ indicating 5.4% of the proportion of variance captured.

The Canonical Loadings in Table 4.3, provided information about the relative contribution of variables to each independent canonical relationship. The first pair of canonical variates can be written as follows:

$X_1 = -\,0.6640\mathit{Mathematics} - 0.4760\mathit{Chemistry} - 0.2566\mathit{Physics}$

$Y_1 = -\ 0.3046Biology - 0.1217Hausa\ Language - 0.3095Economics - 0.5109English\ Language - 0.1941Geography.$

Ø=0.6681

The correlation $Ø_1$ between X_1 and Y_1 is called the first canonical correlation coefficient.

Of the individual variable, Physics loading is the heaviest with the value (-0.2566) followed by Chemistry (-0.4760) and Mathematics (-0.6640) loading for the ordering of the criterion variables.

The values attached to each subjects are their partial correlation to their corresponding canonical variables and indicating the individual contribution to the canonical pair.

Table 4.4 shows the Canonical Cross Loading of the two canonical functions. In the first canonical function, we see that both Chemistry and Physics slightly have high correlations with independent canonical variate-0.4242 and -0.4147 respectively. While the weakest correlation came from Set – 2, i.e. English Language with -0.5463 followed by Geography with -0.4737.

The first canonical correlation explains the maximum relationship between the canonical variates while each successive canonical correlation is estimated so as to be orthogonal and yet explains the maximum relationship not accounted for by the previous canonical correlation. This reflects the high variance among these variables. By squaring the terms in the canonical loading, we find percentage of the variance for each of the variable explained by function 1.

Table 4.5 shows the result of Bartlett's test. The Bartlett's test requires measuring the Homogeneity of variance across variables.From the table, we reject H_0 and conclude that the variances across the variables are not equal. In this regard, this calls for the use of Factor Analysis; to see the variable that poses high variability contribution to the set of data.

Table 4.6 shows the Eigenvalues in column two, which are the proportions of total variance in the entire variable, which are accounted for by the factors. From the output, factor one gives the highest variance explained, followed by factor two, which gives the second highest variance and so on. The second factor is formed from the variance remaining after those associated with the first factor has been extracted, thus this account for the second largest amount of variance.

It is worthwhile to note that the factor coefficient, which gives the variance explained for each factor gives the value of less than 35% of the variance explained.

Therefore, more than one factor is needed to describe the variability of the data. In other to obtain a meaningful interpretation of the Factor Analysis, we need to reduce to fewer than eight (8) factors.

In this study, we use the common decision in which we retain only the factor with about 80% of variance explained. Therefore, from column 3 i.e. extraction sum of squared loading, we observe that five factors are retained together with their percentage of variance explained by each factor. The cumulative variance as given as well shows that the first five factors account for 98.01% of the total variance in the data. A factor's Eigenvalue may be computed as the sum of its squared factor loadings for the entire variable (Rencher, 2002).

A factor's Eigenvalue divided by the number of variables (which equals the sum of variances because the variance of a standardized variance equals to 1.0) gives the percentage of variance in all the variables which it explains. The ration of Eigenvalue is the ratio of explanatory importance of the factor with respect to the variable. So, if a factor has a low

Eigenvalue less than the standardized variance i.e. 1, then it is contributing little to the explanatory importance of variance in the variable and may be ignored as redundant with more important components.

Table 4.7 shows the communalities which measures the percentage of variance explained by all the components. That is, the communality is the squared multiple correlation for the variable using the components as predictors. Communalities for a variable is the sum of squared components loadings for that variable (row) and is the percent of variance due to the variable explained by all the components.

For full orthogonal Factor Analysis, the communality will be 1.0 and all the variance in the variable will be explained by all the factors, with their number equals that of the variables and is written under initial. The extracted communalities, is the percent of variance in a given variable explained by the factors that are extracted, which are normally fewer in number that the original variables which led the coefficient to be less than 1.0

Table 4.8 is going to be used for interpretation. Interpretation for Factor Loadings in Factor Analysis is similar to interpretation of coefficients for Principal Component Analysis and coefficients in multiple regressions as well as Canonical loadings in Canonical Correlation Analysis. We want to have some criterion, which helps us to determine which of these are large and which of these are considered to be negligible.

1. In Factor 1, the percentage amount of variability explained, contributed by the coefficient of each variable is as follow: Hausa Language has the highest coefficient with (-0.0847) followed by Chemistry (-0.1132) and Biology (-0.1331), and to a lesser extent Economics (-0.1654).

2. In Factor 2, Economics with (0.7648) has the highest contribution to the variability, followed by Biology (0.6786), followed by Chemistry (0.5542) and Physics (0.3176).

3. Factor 3 is primarily related to Hausa Language (0.5141) and Mathematics (0.1775).

4. Factor 4 is primarily related to Physics (0.5594) and Geography (0.2522).

5. Factor 5 is primarily related to Chemistry (0.0440) and Biology (-0.0729)

From the Table 4.7 of communalities, it can be seen that all the subjects are well represented. We can think of the value as multiple R^2 values for regression model predicting the variable of interest. The communality for a given variable can be interpreted as the proportion of variance in that variable explained by the five factors. In other word, if multiple regression is performed on Geography against the five factors, therefore R^2 = 0.6351 which is about 63.5% of the variable due to variation in Geography, as explained by factor model. The results suggest that Factor Analysis does the best work of explaining the variation in Geography.

CHAPTER FIVE

Summary, Conclusion and Recommendations

5.0 Introduction

This chapter gives the summary, conclusion and recommendation reached in the canonical correlation analysis and factor analysis techniques used on the entire variables/data.

5.1 Summary from output of Canonical Correlation Analysis

The study aimed at analyzing the Senior Secondary Certificate Examination Results of Command Day Secondary School, Kaduna. The canonical correlation analysis generated three correlation coefficients, which were tested and found one of the correlations statistically different from zero.

From table 4.1, it can be seen that the first canonical pair captured the variability of about 68.2%, the second canonical pair captured the variability of about 26.4% and the third canonical pair captured the variability of about 5.4%. Hence the total variability captured by the three canonical pairs is 100%. The 68.2% variability is due to the individual contribution of the composites of Mathematics, Physics, Chemistry, Economics, English Language, Biology, Geography and Hausa Language, showing that, there is slight relationship among the above listed subjects.

5.2 Summary from Output of Factor Analysis

Factor Analysis was run on the whole data. Based on the decision of any factors that cover a certain percentage say 80% and the scree plot test, five factors were considered. The first, second, third, fourth and fifth factors captured the variability 28.73%, 33.89%, 8.96%, 11.43% and 15.01% respectively. The total variability captured due to the four factors is 83.01%. Each factor represents a group of closely related subjects,

The contribution of the first group is due to all, however, Hausa language, Chemistry and Biology contributed significantly. The second is due to the contribution of Economics, Biology and Chemistry. The third is due to the contribution of Hausa Language and Mathematics. The fourth is due to Physics and Geography, while finally, the fifth is due to Chemistry and Biology. However, those values under Table 4.8 that are very close to zero were dropped, which resulted to data reduction.

5.3 Conclusion

It is clearly shown that set-1 and set-2 are correlated as sought for; Canonical Correlation analysis measured the strength of relationship of the canonical pair and the subjects that strongly contributed. The first pair with a measure of correlation of 0.6681 with the proportion of variability of about 68%, the second pair with a measure of correlation 0.4152 with the proportion of variability of about 27% and the third canonical pair with a measure of correlation 0.1883 having a proportion of variability of about 5%.

Factor analysis was also applied and showed five groups of closely inter-related subjects based on the fact that five factors were used. It is also shown in Table 4.8 that values that is close to zero correlating a variable were dropped, which indicates variable reduction. The strongest inter-related subjects are found in the beginning column of table 4.8 and decrease through the last column. The 28.73% of the variability captured by the inter-related variables is due to the contribution of all the subjects except Mathematics, Physics and partially Chemistry.

Mathematics, Physics and Chemistry are directly related, whereas the other subjects are also related to each other. However, it can be seen that performance of students in Mathematics, Physics and Chemistry is inversely related to the other subjects. That is, an increase in the performance in Mathematics, Physics and Chemistry will lower the performance of the other subjects and this might be due to much emphasis and concentration given to the mathematical subjects.

5.4 Recommendations

From the output of the analysis carried out on the entire data, it is clearly seen that the correlation between the mathematical and less-mathematical subjects is averagely moderate at the first canonical variate due to much emphasis and concentration given to the mathematical subjects compared to the less-mathematical by the authority of the school. Better understanding of less-mathematical subjects will enhance good understanding of the mathematical subjects as all its expressions and explanation contains English Language.

5.5 Contribution to Knowledge

1. In this thesis we have shown that Mathematical subjects i.e. Mathematics, Physics and Chemistry are directly related, so also the less mathematical subjects i.e. English Language, Biology, Economics, Geography and Hausa Language. Moreover, it was shown that the mathematical subjects are inversely related to less mathematical subjects.

2. This is the first time as far as we know that Canonical Correlation and Factor Analysis are combined in determining the performance of students in their academic pursuits.

REFERENCES

Antonia L. D. (2003). Personal, family, and Academic Factors Affecting Low Achievement in Secondary School.*Electronic Journal of Research in Education Psychology and Psychopedagogy, 1 (1) 43 – 66.* ISSN: 1696 - 2095.

Affendey, L.S., Mustapha, N., Nasir, S., and Muda, Z. (2010).Ranking of Influencing Factors in Predicting Students' Academic Performance.*Information Technology Journal 9 (4): 832-837*.

Anastasi, A. and Urbina, S. (1997), *Psychological Testing upper saddle River* N.J Prentice Hall.

Anderson, T.W. (1958). *An Introduction to Multivariate Statistical Analysis,* First Edition, John Wiley and Sons, New York.

Ballester, M.A., Linguraru, M.G., Aguirre M.R., and Ayache, N.,(2004).*"On the Adequacy Variability INRIA,* Epidaure Research Project 2004 route deslucioles BP3, Sophia Antipolis, 06902, France http://www.inrla,fr/epidaure.

Bartholomew, D.J., Steele, F., Galbraith, J., andMoustaki, I. (2008). *Analysis of Multivariate Social Science Data*. Statistics in the Social and Behavioral Sciences Series (2nd ed.). Taylor & Francis. ISBN 1584889608.

Bartlett, S., Peel, M.J. and Pendlebury M. (1993).From fresher to finalist: a three year analysis of student performance on an accounting degree programme, *Accounting Education 2(2), 111-122.*

Bartlett, M.S. (1937). The Statistical Significance of Canonical Correlations,*Biometrica 32:29.*

Bonaci C. G., Mutju, A. and Mustata R. V.(2010).Influential Factors of Accounting Students' Academic Performance: A Romanian Case Study, *Accounting and Management Information Systems 9(4), 558-580*.

Borga, M. (2001).*Canonical Correlation a Tutorial,* http://people.imt.liu.se/~magnus/cca. Unpublished

Chamillard, A.T., (2006). Using Student Performance Predictions in a Computer Science Curriculum.*Proceeding of the 11th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education, June 26-28, Bologna, Italy, 260-264*.

Collier, H. and Mehrens, W. (1985).Using multiple choice test items to improve classroom testing of professional accounting students, *Journal of Accounting Education 3(2), 41-51*.

David, W. (2003).Canonical Correlation Analysis, *Proceeding 25, Institute of Phonetic Sciences, University of Amsterdam 81-99.*

De Zoysa, A. and Kathy, R. (2006). Implications for Accounting Educators of Student Socio-Economic Circumstances,*Accounting & Finance Working Paper 06/24,* School of Accounting & Finance, University of Wollongong.

De la Fuente, J. (2002). Recent Perspectives on the Study of Motivation; the Theory of Goal Orientation.*Escritos de Psicologia, 2, 72 – 84.*

Everitt, B.S. and Dunn, G. (1991).*Applied Multivariate Data Analysis,* Edward Arnoid, London, 219 – 220.

Golding, P. and Donaldson, O. (2006). Predicting Academic Performance.*Proceedings of the 36th ASEE/IEEE Frontiers in Education Conference T1D-21, Oct. 28-31,* San Diego, C.A., 1-6.

Gose, B. (1998). Tutoring Companies Take over Remedial Teaching at Some Colleges: Can Kaplan and Sylvan Help Students Erase Educational Deficiencies More Quickly? *The Chronicle of Higher Education, 35 (12).*

Guney, Y. (2009). Exogenous and Endogenous Factors Influencing Students' Performance in Undergraduate Accounting Modules.*Accounting Education, 18, (1), 51-73*.

Hardle, W. and Simar, L. (2003).*Applied Multivariate Statistical Analysis,* version 29th April, 2003, Berlin.

Harlow, L. (2005). *The Essence of Multivariate Thinking, Theme and Methods,* Mahwah N.J. Lawrence Erlbaum Associates, 205, 229.

Hotelling, H. (1936). Relations between Two Sets of Variables, *Biometrika, 28:312 - 377*

Howard, T. and Diane, T. (1987).Uses of Factor Analysis in Counseling Psychology Research, *Journal of Counseling Psychology 34 (4) 414 – 424*.

Jordan, J. and Suzanne, G.S. (2006).The Effect of Moral Orientation Variables in Counseling College Students, *Journal of College of Character7, 7.*

Koh, M. Y. and Koh H.C. (1999).The Determinants of Performance in an Accountancy DegreeProgramme*Accounting Education 8(1), 13-29*.

Liu, T. (2004). Predict Modeling of Atomic Structure Using Canonical Correlation Analysis, University of Pennsylvania, *Philadephia*.

McKenzie, K. and R. Schweitzer, (2001).Who succeeds at University? Factors predicting academic performance in first year Australian university students.*Higher Educ. Res. Dev., 20: 21-33.*

Montero, M.C. (1990).Predicting Academic Performance.A Study of Intervening Variables in a Sample of 8th Grade Students with Follow-up in 10th Grade.Tesis.Universidad Pontificiade Salamanca.

Mustata R. V., Bonaci C. G. and Mutju A, (2010). Determinants of Accounting Students' Academic Performance: A Factor Analysis Based Approach.*Accounting and Management Information Systems 9 (4), 580-600.*

Naser, K. and Peel, M. (1998).An exploratory study of the impact of intervening variables on student performance in a Principles of Accounting course.*Accounting Education, 7(3): 209-223*.

Nonis, S. and Hudson, G. (2006), "Academic performance of college students; influence of time spentstudying and working" *Journal of Education for Business 81(3), 151- 160*.

Ranjan, R. and Nallari, R. (2005).Study of Asset Liability Management in Indian Banks, Canonical Correlation Analysis.*Accounting and Management Information Systems 9 (4), 450-461*.

Rencher, A.C. (2002). Methods of Multivariate Analysis, 2nd Edition, Published by John Wiley and Sons.

Rencher, A.C. (1998). On the Use of Correlation to Interpret Canonical Functions, *Biometrika, 75, 363 – 365.*

Richard, A. and Dean, W. (1992).*Applied Multivariate Statistical Analysis.* Prentice Hall, Englewood Cliffs, New Jersey 07632.

Shafto, M.G. (1997). Canonical Correlation Analysis of Data on Human Automation Interaction, *Proceeding of the 41st Annual Meeting of the Human Factors and Ergnomics Society*.

Shen, H., Jegelka, S., and Gretton, A. (2009) Fast Kernel-based Independent Component Analysis.*IEEE Transactions on Signal Processing, 57(9):3498.*

Simon, H.A (1969).*The Science of the Artificial,* Cambridge M.A., M.I.T Press.

Slater, J.N. (2002). Application or Motivation Theory; An Analysis of the Motivation of At-risk Ninth Grade Students Enrolled in Online Courses. *Dissertation Abstracts International, Section A: Humanities and Social Sciences, 62.*

Snedecor, G. W. and Cochran, W. G. (1989).*Statistical Methods,* Eighth Edition, Iowa State University Press.

Superby, J.F., Vandamme J.P. andMeskens,N. (2006). Determination of Factors Influencing the Achievement of the First-year University Students Using Data Mining Methods.*Proceedings of the 8th International Conference on Intelligent Tutoring Systems, Educational Data Mining Workshop, (ITS`06), Jhongali, Taiwan, 37-44*.

Tapia, A. (2002). Personal, Family and Academic Factors Affecting Low Achievement in Students Performance.*Electronic Journal of Research in Education Psychology and Psychopedagogy, 1 (1) 43 – 66.* ISSN: 1696 - 2095.

Valle A. J. (1999).Causal Attribution, Self-Concept and Motivation in Students with High and Low Academic Performance.*Revista Espanola de Pedagogia, (214), 525 – 546.*

Vandamme, J.,Meskens, N.andSuperby, J.F.(2007).Predicting Academic Performance by Data Mining Methods.*Educ. Econ. 15: 405-419.*

Wheatley, M.J. (1994). Leadership and the New Science about Organization from an Orderly Universe, San Francisco CA. BerttKoehley Publisher, Inc.

Wooten, T. (1998) "Factors Influencing Student Learning in Introductory Accounting Classes: AComparison of Traditional and Nontraditional Students" *Issues in Accounting Education13, (2), 357-373.*

Yi Chia, H. (2002). The Relationship between Task Value, Self-Efficacy and Students Performance in Performance Assessment.*Dissertation Abstracts International Section A: Humanities and Social Sciences, 62.*