

**MODELLING SCHOOL FACTORS AND PERFORMANCE
IN MATHEMATICS AND SCIENCE IN KENYAN
SECONDARY SCHOOLS USING CANONICAL
CORRELATION ANALYSIS**

BY

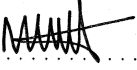
JEREMIAH MBARIA MUCUNU

**A research project submitted in partial fulfillment of the
requirements for the award of the degree of Master of Science
in Social Statistics, School of Mathematics, University of
Nairobi.**

DECEMBER, 2017

Declaration

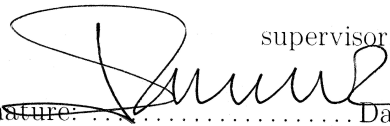
I declare that this is my original work and has not been presented for any award of a degree in any other university.

Signature:  Date: 8/12/2017

Jeremiah Mbaria Mucunu

I56/69527/2013

This research project has been submitted for examination with my approval as

 supervisor.
Signature: Date: 8-12-2017

Dr. George Muhua

DEDICATION

I dedicate this work to my parents Prof. James Mbaria and Mrs. Anneh Mbaria and to my wife Valentine Wairimu and our daughter Arianna Wanjiru.

ACKNOWLEDGEMENT

I would like to express my deep gratitude to my supervisor Dr. George Muhua for the great supervision and guidance. My sincere appreciation also goes to Prof. Patrick Weke, Director School of Mathematics and all the lecturers who assisted me in my course work.

Words can not express how grateful I am to my parents for their tremendous support throughout this journey. I must express my very profound gratitude to my parents in law, Duncan Kariuki and Monicah Kariuki. Your prayers and encouragement kept me going. I owe a debt of gratitude to my wife for creating the best environment for me to work. Special mention goes to Loise, Josiah, George, Benard, Stella, Margaret and Imani for their motivation and inspiration.

Finally, I thank the Almighty God for His sufficient grace that has made this possible.

ABSTRACT

The level of performance and participation in Science, Technology, Engineering and Mathematics (STEM) career subjects remains low in Kenya despite STEM's critical role in economic development. Numerous factors contribute to students' academic achievement in STEM education. This study focusses on modelling school factors that affect the performance in mathematics and science in Kenyan secondary schools using Canonical Correlation Analysis (CCA). The objectives of the study include determining: the magnitude of the relationship between school factors and performance in STEM education, the most influential subject in describing the level of STEM education, the most contributing school factor to STEM education and a model to predict performance in STEM education given school factors. This research utilised data from 9,834 candidates of year 2015 Kenya Certificate of Secondary Education (KCSE) from 77 public secondary schools in Nairobi County. CCA is a multivariate data analysis technique that seeks to establish whether two sets of variables, predictor and criterion, are independent of each other. Given that the two sets of variables are dependent, CCA is able to represent a relationship between the sets of variables rather than individual variables. From the 2015 KCSE data, CCA revealed that school factors significantly correlate with the level of performance in STEM education. Based on standardised canonical coefficients and canonical loadings, the subjects that mainly influence the level of performance in STEM education were found to be mathematics and physics. Further assessment of the canonical cross loadings from the two variate pairs revealed that the proportion of students with mean grades of C+ and above and the proportions of students taking biology and physics were found to contribute very highly to the level of performance in STEM education. The study recommends increased staffing in physics due to the fact that physics is an optional subject yet it has comparatively larger loadings than biology and chemistry which have higher levels of participation. Also, the study recommends that further studies should be done to establish the relationship between individual factors and participation and performance in STEM career subjects.

TABLE OF CONTENTS

DECLARATION	i
DEDICATION	ii
ACKNOWLEDGEMENT	iii
ABSTRACT	iv
LIST OF ABBREVIATIONS	viii
LIST OF FIGURES	ix
LIST OF TABLES	x
1 INTRODUCTION	1
1.1 Background of the study	1
1.2 Statement of the problem	3
1.3 Objectives of the study	3
1.3.1 Main objective	3
1.3.2 Specific objectives	3
1.4 Significance of the study	4
2 LITERATURE REVIEW	5
2.1 Introduction	5
2.2 School factors and academic performance	5
2.3 Summary of literature review and knowledge gaps	7
3 METHODOLOGY	8
3.1 Introduction	8
3.2 Source of Data	8
3.3 Definition of variables	8

3.4	Canonical Correlation Analysis	9
3.5	Aims of Canonical Correlation Analysis	12
3.6	Scope of Canonical Correlation Analysis	12
3.7	Canonical variate pairs	12
3.8	Canonical correlation coefficients	13
3.8.1	Deriving the first canonical variate pair	13
3.8.2	Deriving the second canonical variate pair	16
3.8.3	Deriving the i^{th} canonical variate pair	17
3.9	Canonical weights	18
3.10	Standardized coefficients	18
3.11	Canonical loadings	19
3.12	Canonical cross loadings	19
3.13	Canonical variate scores	20
3.14	Tests of significance	20
3.14.1	Tests of independence between X - set and Y - set	20
3.14.2	Tests of significance of the i^{th} variate	22
3.15	Assumptions of Canonical Correlation Analysis	23
4	DATA ANALYSIS, RESULTS AND DISCUSSIONS	24
4.1	Introduction	24
4.2	Exploratory Data Analysis	24
4.3	Canonical Correlation Analysis	27
4.3.1	Correlations	27
4.3.2	Test of independence between X - set and Y - set	28
4.3.3	Test of significance of the second variate pair	29
4.3.4	Eigenvalues and Canonical Correlations	29
4.3.5	Canonical weights	30
4.3.6	Canonical loadings	31
4.3.7	Canonical cross loadings	33

4.3.8 Prediction	33
5 SUMMARY OF FINDINGS, CONCLUSIONS AND RECOMMENDATIONS	36
5.1 Summary of Findings	36
5.2 Conclusions	36
5.3 Recommendations	37
REFERENCES	38
APPENDICES	40
A CAREER CLUSTERS AND PATHWAYS	41
B KCSE 2010 TO 2015 NATIONAL PERFORMANCE	43

LIST OF ABBREVIATIONS

ANOVA	Analysis of variance
ATPS	African Technology Policy Studies Network
CCA	Canonical Correlation Analysis
EDA	Exploratory Data Analysis
INSET	In-service Education and Training
KCSE	Kenya Certificate of Secondary Education
KNEC	Kenya National Examination Council
MOEST	Ministry Of Education Science and Technology
STEM	Science, Technology, Engineering and Mathematics
ST&I	Science Technology and Innovation
SMASSE	Strengthening of Mathematics and Science in Secondary Education
SMT	Science, mathematics and technology
UNESCO	United Nations Educational, Scientific and Cultural Organization
TSC	Teachers Service Commission

LIST OF FIGURES

4.1	Scatter plot for the criterion set of variables	26
4.2	Scatter plot for the predictor set of variables	27

LIST OF TABLES

4.1	Summary statistics for KCSE examination results in STEM subjects and school characteristics data	24
4.2	Correlations within and between the predictor and criterion two sets of variables	28
4.3	Multivariate Tests of Significance ($S = 4, M = 0, N = 33$)	28
4.4	Dimension Reduction Analysis	29
4.5	Eigenvalues and Canonical Correlations	29
4.6	Raw and Standardized canonical weights	30
4.7	Canonical Loadings	32
4.8	Canonical Cross Loadings	33

CHAPTER 1

INTRODUCTION

1.1 Background of the study

Education plays a key role in the professional career development of an individual. Schools influence students' values, attitudes and career selection. In collaboration with parents, guardians and employers, teachers prepare students to take numerous roles that they choose to be engaged in during their lives. Thus, one purpose of education is preparing students for employment. It is therefore important that schools are well equipped to help children in career development. According to Patton and McMahon (2014) career development takes place in five phases based on the child's age. The five phases are associated with tasks that assist students in career decision making at various age brackets. Students in secondary schools are in the exploration phase. At this phase, they explore various occupational clusters. Through this exploration they acquire an initial work experience.

Each and every career falls into one of the sixteen career clusters developed by the States' Career Clusters Initiative in 2002 (Carnevale et al., 2011). A career cluster is defined as a set of occupations and activities that relate to each other by the types of products and skills. Every cluster corresponds to an array of courses that prepare students for a given career referred to as career pathways. Appendix A contains the list of the sixteen career clusters and their corresponding career pathways. This study focusses on the Science, Technology, Engineering and Mathematics (STEM) career cluster. STEM education refers to the teaching and learning of science, technology, engineering, and mathematics (Gonzalez and Kuenzi, 2012). The gender gap in terms of participation in STEM careers has been narrowing over the years. It has been documented that compared to men, members of the female gender who embark on a career in STEM later leave their jobs to concentrate on family engagements (Huyer,

2015).

Kenya, amongst many other countries, is widely believed to perform poorly in STEM education. The performance of secondary school students in science and mathematics has been very poor, compared to other subjects, between 2010 and 2015. Appendix B shows the overall subject means for various subjects in the Kenya Certificate of Secondary Education (KCSE) examinations. Due to the low level of performance and participation in STEM career subjects at secondary school level, few students pursue related courses at the university level. Data from the Ministry of Education Science and Technology (MOEST) in Kenya reveals that about 22% of students in universities in the year 2016 were enrolled for courses in STEM. The rest (78%) were in humanities and social sciences. To address the current shortages and deficiencies in STEM education there is need to prepare and equip teachers adequately (The World Bank, 2016).

Kenya recognizes the importance of STEM in the realization of its vision 2030. The Government adopted the National Science, Technology and Innovation (ST&I) Policy and Strategy. This direct and promotes the absorption of ST&I in all sectors of the economy. In an effort to improve STEM education in Kenya, the government has also institutionalized In-service Education and Training (INSET) sessions for teachers who teach mathematics and sciences under the Strengthening of Mathematics and Science in Secondary Education (SMASSE) programme. The United Nations Educational, Scientific and Cultural Organization (UNESCO) emphasises the importance of ensuring that curriculum is sensitive to gender with regards to STEM education so as to achieve Kenya Vision 2030 (Nagel, 2017). The characteristics of a student's former secondary school have a greater impact on the academic performance of that student at the university than the student's own background characteristics (Win and Miller, 2005). Hence, a country's schools and academic system play a vital role in influencing students' interest in STEM subjects. In school, students get equal opportunities to participate and perform well in STEM education (Bryant, 2012).

1.2 Statement of the problem

Performance and participation in STEM career subjects remains low in Kenya despite STEM's critical role in economic development. Numerous factors contribute to students' academic achievement in STEM education. There is a need to focus on the contribution schools make in assisting students to achieve better scores in STEM. Due to the challenges of poor performance in STEM career subjects, this study endeavors to establish the relationship between the school characteristics and academic achievement in STEM education.

1.3 Objectives of the study

1.3.1 Main objective

To establish the relationship between Kenyan secondary schools' characteristics and students' level of academic performance in science and mathematics.

1.3.2 Specific objectives

- (i) To determine the most contributing subject in defining the level of performance in science and mathematics
- (ii) To determine the most influential variable in defining school characteristics with regard to STEM education
- (iii) To determine the magnitude of the relationship that exists between school characteristics and performance in science and mathematics
- (iv) To predict the performance in science and mathematics given the school characteristics

1.4 Significance of the study

While numerous studies have studied determinants of performance in science and mathematics using explorative and descriptive statistics, such studies do not specifically consider how individual STEM career subjects are influenced by school characteristics. Therefore, the approach taken in this project is to use canonical correlation analysis (CCA) rather than other statistical methods used on the previous studies to establish the relationship between performance in STEM subjects and school characteristics. Furthermore, if this relationship is established, it will help provide important information to the MOEST and other stakeholders in policy and strategy formulation to facilitate improved participation and performance in STEM education.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

The following literature review confirms that schools have a major influence on students' participation and performance in science and mathematics subjects. It highlights different approaches taken to establish the magnitude of influence schools have on students' academic success, which highly impacts on students' career development.

2.2 School factors and academic performance

A study conducted by Albert, Osman, and Yungungu (2014) investigated the factors that influence performance in Biology in the Kenya Certificate of Secondary Education (KCSE) examination. The sample constituted 730 students, 18 biology teachers and 14 principals from 14 selected schools in Nyakach District, Kisumu County. Data was obtained from the sample using interviews and questionnaires. The dependent variable in the study was performance in Biology and the independent variables included teacher characteristics, availability of teaching and learning resources, motivation and students' attitude towards Biology. Separate correlation analyses were conducted on each independent variable and the results show that there was positive relationship between performance in KCSE Biology and teacher characteristics, teaching and learning resources, motivation and students' attitude towards Biology. The highest positive correlation was between performance in KCSE biology and teacher characteristics.

Mbaki, Joash, and Muola (2010) studied the factors affecting girls' performance in science, mathematics and technology (SMT) in public secondary schools in Kenya. This study looked into the effect of school factors on academic performance in mathematics, chemistry, biology, physics and agriculture. The school factors included teacher qualification,

teaching load, availability of teaching and learning resources and class size. The data was obtained from 30 SMT teachers, 6 head teachers, 416 girls who participated in KCSE in the year 2009 from 6 secondary schools in Kitui Central District. The performance in SMT was represented by the average scores in SMT subjects which were categorized into three levels: below average, average and above average. Analysis of variance (ANOVA) tests were performed on each of the variables. It was found that there were statistically significant differences among the three academic levels in terms of teachers' teaching load, availability of teaching and learning resources and class size. However, there was no significant difference between teacher qualifications and girls' performance in SMT subjects. Mbaki, Joash, and Muola (2010) performed correlation analyses to explore the differences revealed by the ANOVA tests. Based on the sample used, performance in SMT subjects was improved by smaller teaching loads, more availability of teaching and learning resources and smaller class sizes.

Win and Miller (2005) studied how University students' academic performance is affected by individual and school factors. This research was conducted on 1,803 first year students who entered the University of Western Australia in 2001. The students' first year academic performance was the dependent variable. The explanatory variables constituting the individual factors included the students' prior academic achievement at high school level, the gender, the home location, the economic status and the education level at home. The explanatory variables constituting the school factors included the type of high school attended, the proportion of graduates from that school and the percentage of students who passed the entry examinations from that school. Win and Miller (2005) used a linear regression model to analyse the data. Since the explanatory variables were in two levels, the students were nested within schools. This way, individual level variables are separate from the school level variables within the model. This study revealed that the previous secondary school has the greatest impact on the academic performance of students at the university compared to the background characteristics of students'.

2.3 Summary of literature review and knowledge gaps

Albert, Osman, and Yungungu (2014) focuses on only one STEM career subject, biology. It is of interest to determine how other science and mathematics subjects are influenced by the stated factors. Mbaki, Joash, and Muola (2010) explore the impact of school factors on several STEM career subjects providing more information about STEM education. However, the contributions of teacher characteristics to STEM education from these two studies are contradictory. This could be attributed to the fact that Mbaki, Joash, and Muola (2010) use an average score of several subject means, which is not an accurate measure. Some subjects used in the calculation of the mean score are optional, implying that the measure is unweighted. Hence, there is need to consider the effects of such differences. Use of weighted means gives more consistent estimates (Solon, Haider, and Wooldridge, 2015).

Win and Miller (2005) succeed in highlighting the most contributing set of variables in predicting students' academic performance. However, when hierarchical data is dealt with on a one-level basis problems of aggregation bias arise. Other related problems include multicollinearity, failure to satisfy the assumptions of independence and heterogeneity of regression (Bickel, 2007).

It is important to determine the influence of a set of factors on a set of STEM career subjects. Canonical correlation analysis (CCA) is suitable in filling the identified knowledge gaps. CCA reduces the probability of having Type I errors. When several tests in statistics are applied for each dependent variable, the probability of making a Type I error increases (Thompson, 1991).

CHAPTER 3

METHODOLOGY

3.1 Introduction

This chapter highlights the sources of data and the definition of variables used in the study. The objectives, scope, assumptions and derivations of the canonical correlation analysis (CCA) model are also presented in this chapter.

3.2 Source of Data

The data used in this research is obtained from the Kenya National Examination Council (KNEC) and the Teachers Service Commission (TSC). The data comprises of the Kenya Certificate of Secondary Examination (KCSE) results of the year 2015 for 9,834 candidates from 77 public secondary schools in Nairobi County.

3.3 Definition of variables

The overall objective of this study is to establish the relationship between school characteristics and performance in science and mathematics subjects. From the KCSE results data, two sets of variables can be generated. The independent set of variables contains the school characteristics which are defined by the teacher to student ratio, school size, percentage of students taking biology, percentage of students taking physics and percentage of students with mean grades above C+. The dependent set of variables represents performance in science and mathematics and is given by mean scores in mathematics, biology, physics and chemistry.

3.4 Canonical Correlation Analysis

Canonical correlation analysis (CCA) is a multivariate statistical technique that enables the establishment of linear relationships between two groups of variables, independent and dependent variables. This study seeks to establish if five X variables, school characteristics, can predict four Y variables, performance in STEM career subjects. This is different from multiple regression, where separate relationships are obtained for each dependent variable. CCA is able to denote a relationship between a group of variables rather than single variables. Moreover, it can identify unique relationships in two or more levels, if they exist (Thompson, 1991).

A canonical variate is created for each group of variables. A canonical variate is the linear combination obtained from the group of independent variables in a multiple regression analysis. In CCA there is an additional variate obtained from several dependent variables. Consider two groups of variables, X and Y. Suppose the number of variables for X and Y are q and p respectively. The data vectors of X-set and Y-set can be written as:

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_q \end{pmatrix} \quad (3.1)$$

and

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{pmatrix} \quad (3.2)$$

The linear combinations $\mathbf{X}^* = \mathbf{a}'\mathbf{X}$ of the variables in the X-set and $\mathbf{Y}^* = \mathbf{b}'\mathbf{Y}$ of the variables in the Y-set, where \mathbf{a} and \mathbf{b} are two vectors of constants of elements q and p respectively, are referred to as *canonical variates*. The canonical variates are be

expressed as

$$X^* = a'X = \begin{pmatrix} a_1 & a_2 & \dots & a_q \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_q \end{pmatrix} = a_1X_1 + a_2X_2 + \dots + a_qX_q \quad (3.3)$$

and

$$Y^* = b'Y = \begin{pmatrix} b_1 & b_2 & \dots & b_p \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{pmatrix} = b_1Y_1 + b_2Y_2 + \dots + b_pY_p. \quad (3.4)$$

The variables X^* and Y^* are called ***canonical variables***. The coefficients of X and Y in the linear composites are called ***canonical weights or coefficients***. The sample observations from the data vectors in equation 4.1 and equation 4.2 can be augmented as

$$\begin{pmatrix} \mathbf{x}_i \\ \mathbf{y}_i \end{pmatrix} = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iq} \\ y_{i1} \\ y_{i2} \\ \vdots \\ y_{ip} \end{pmatrix}, i = 1, 2, \dots, n. \quad (3.5)$$

For the sample of size n , the mean vector is given by equation 3.6. The covariance matrix and correlation matrix are represented by equation 3.7 and equation 3.8 respectively.

$$\begin{pmatrix} \bar{\mathbf{x}} \\ \bar{\mathbf{y}} \end{pmatrix} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_q \\ \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_p \end{pmatrix}, \quad (3.6)$$

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{xx} & \mathbf{S}_{xy} \\ \mathbf{S}_{yx} & \mathbf{S}_{yy} \end{pmatrix} = \begin{pmatrix} \mathbf{S}_{xx} & \mathbf{S}_{xy} \\ \mathbf{S}'_{xy} & \mathbf{S}_{yy} \end{pmatrix} \quad (3.7)$$

and

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_{xx} & \mathbf{R}_{xy} \\ \mathbf{R}_{yx} & \mathbf{R}_{yy} \end{pmatrix} = \begin{pmatrix} \mathbf{R}_{xx} & \mathbf{R}_{xy} \\ \mathbf{R}'_{xy} & \mathbf{R}_{yy} \end{pmatrix} \quad (3.8)$$

respectively. For the covariance matrix, \mathbf{S}_{xx} is $q \times q$, \mathbf{S}_{xy} is $q \times p$, \mathbf{S}_{yx} is $p \times q$ and \mathbf{S}_{yy} is $p \times p$. Similarly, for the correlation matrix, \mathbf{R}_{xx} is $q \times q$, \mathbf{R}_{xy} is $q \times p$, \mathbf{R}_{yx} is $p \times q$ and \mathbf{R}_{yy} is $p \times p$. The analogous population results for equation 3.6 and equation 3.7 are

$$E \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} E(\mathbf{x}) \\ E(\mathbf{y}) \end{pmatrix} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \text{ and} \quad (3.9)$$

$$cov \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix} = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma'_{xy} & \Sigma_{yy} \end{pmatrix}. \quad (3.10)$$

If there is no linear relationship between the groups, \mathbf{x} and \mathbf{y} , then $\Sigma_{xy} = \mathbf{O}$.

3.5 Aims of Canonical Correlation Analysis

Objectives of CCA are to;

- (i) Determine the measure of the linear relationships between two sets of variables
- (ii) Derive coefficients for each group of variables in order to maximise the correlation between the two groups
- (iii) Explain the type of relationships that exists between two groups of variables

3.6 Scope of Canonical Correlation Analysis

CCA studies linear relationships between two groups of variables. In addition, CCA is appropriate when two groups of variables are measured on each sampling unit. CCA can be applied on both metric and non-metric data. CCA is appropriate when there exists correlation between dependent variables (Bhuyan, 2005).

3.7 Canonical variate pairs

CCA provides canonical variate pairs, where a variate in a pair is either a linear composite of variables in X-set or a linear composite of variables in Y-set. Given that the number of variables in the groups X and Y are q and p respectively, the maximum number of pairs is $k = \min(p, q)$. Variate pairs are selected such that each pair is highly correlated and subsequent pairs are independent of each other. The i^{th} canonical variate pair is defined by (X_i^*, Y_i^*) . Thus, the first canonical variate pair is given by (X_1^*, Y_1^*) .

3.8 Canonical correlation coefficients

The canonical correlation coefficient for the i^{th} pair of variates is the correlation between X_i^* and Y_i^* and is calculated using the formula:

$$\rho_i^* = \frac{cov(X_i^*, Y_i^*)}{\sqrt{var(X_i^*)var(Y_i^*)}} \quad (3.11)$$

CCA formulates an equation linking the X and Y variables that maximizes the canonical correlation coefficient between the pair of variates.

3.8.1 Deriving the first canonical variate pair

First canonical variate pair can be denoted as

$$\begin{aligned} \begin{pmatrix} X_1^* \\ Y_1^* \end{pmatrix} &= \begin{pmatrix} a_{11}X_1 + a_{12}X_2 + \dots + a_{1q}X_q \\ b_{11}Y_1 + b_{12}Y_2 + \dots + b_{1q}Y_q \end{pmatrix} = \begin{pmatrix} a_1'X \\ b_1'Y \end{pmatrix} \\ &= \begin{pmatrix} a_1' & \mathbf{O}' \\ \mathbf{O}' & b_1' \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = \mathbf{AB}. \end{aligned} \quad (3.12)$$

Thus $\begin{pmatrix} X_1^* \\ Y_1^* \end{pmatrix}$ has a population covariance matrix

$$\begin{aligned} A\Sigma A' &= \begin{pmatrix} a_1' & \mathbf{O}' \\ \mathbf{O}' & b_1' \end{pmatrix} \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix} \begin{pmatrix} a_1' & \mathbf{O}' \\ \mathbf{O}' & b_1' \end{pmatrix} \\ &= \begin{pmatrix} a_1'\Sigma_{xx}a_1 & a_1'\Sigma_{xy}b_1 \\ b_1'\Sigma_{yx}a_1 & b_1'\Sigma_{yy}b_1 \end{pmatrix} \end{aligned} \quad (3.13)$$

hence the first population canonical correlation coefficient is given by

$$\rho_1^* = \rho_{X_1^*Y_1^*} = \frac{a_1'\Sigma_{xy}b_1}{\sqrt{a_1'\Sigma_{xx}a_1}\sqrt{b_1'\Sigma_{yy}b_1}}. \quad (3.14)$$

The sample estimate of equation 3.14 is given by

$$\hat{\rho}_1^* = \hat{\rho}_{X_1^* Y_1^*} = \frac{\hat{a}'_1 \mathbf{S}_{xy} \hat{b}_1}{\sqrt{\hat{a}'_1 \mathbf{S}_{xx} \hat{a}_1} \sqrt{\hat{b}'_1 \mathbf{S}_{yy} \hat{b}_1}}. \quad (3.15)$$

The vectors a_1 and b_1 are selected such that first canonical correlation coefficient is at a maximum. ρ_1^* is maximum when $(\rho_1^*)^2$ is maximum. Letting $V = (\rho_1^*)^2$, we obtain

$$V = \frac{(a'_1 \Sigma_{xy} b_1)^2}{(a'_1 \Sigma_{xx} a_1)(b'_1 \Sigma_{yy} b_1)}. \quad (3.16)$$

V is maximum when the partial derivatives $\partial V / \partial a_1$ and $\partial V / \partial b_1$ are equal to zero. Computing the partial derivative of V with respect to a_1 , we have

$$\frac{\partial V}{\partial a_1} = \frac{1}{(b'_1 \Sigma_{yy} b_1)} \frac{2(a'_1 \Sigma_{xy} b_1) \Sigma_{xy} b_1 (a'_1 \Sigma_{xx} a_1) - 2(a'_1 \Sigma_{xy} b_1)^2 \Sigma_{xx} a_1}{(a'_1 \Sigma_{xx} a_1)^2} = 0. \quad (3.17)$$

Simplifying equation 3.17 gives

$$2(a'_1 \Sigma_{xy} b_1)(a'_1 \Sigma_{xx} a_1) \Sigma_{xy} b_1 = 2(a'_1 \Sigma_{xy} b_1)^2 \Sigma_{xx} a_1$$

$$\Sigma_{xy} b_1 (a'_1 \Sigma_{xx} a_1) = (a'_1 \Sigma_{xy} b_1) \Sigma_{xx} a_1 \quad (3.18)$$

$$a_1 = \frac{(a'_1 \Sigma_{xx} a_1)}{(a'_1 \Sigma_{xy} b_1)} \Sigma_{xx}^{-1} \Sigma_{xy} b_1. \quad (3.19)$$

Similarly, computing the partial derivative of V with respect to b_1 , we have

$$\frac{\partial V}{\partial b_1} = \frac{1}{(a'_1 \Sigma_{xx} a_1)} \frac{2(a'_1 \Sigma_{xy} b_1) \Sigma'_{xy} a_1 (b'_1 \Sigma_{yy} b_1) - 2(a'_1 \Sigma_{xy} b_1)^2 \Sigma_{yy} b_1}{(b'_1 \Sigma_{yy} b_1)^2} = 0. \quad (3.20)$$

Simplifying equation 3.20 gives

$$2(a'_1 \Sigma_{xy} b_1)(b'_1 \Sigma_{yy} b_1) \Sigma'_{xy} a_1 = 2(a'_1 \Sigma_{xy} b_1)^2 \Sigma_{yy} b_1$$

$$\Sigma'_{xy} a_1 (b'_1 \Sigma_{yy} b_1) = (a'_1 \Sigma_{xy} b_1) \Sigma_{yy} b_1 \quad (3.21)$$

$$b_1 = \frac{(b'_1 \Sigma_{yy} b_1)}{(a'_1 \Sigma_{xy} b_1)} \Sigma_{yy}^{-1} \Sigma'_{xy} a_1. \quad (3.22)$$

Substituting equation 3.22 into equation 3.18 we obtain

$$\begin{aligned} \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma'_{xy} a_1 &= \frac{(a_1 \Sigma_{xy} b_1)^2}{(a'_1 \Sigma_{xx} a_1)(b'_1 \Sigma_{yy} b_1)} \Sigma_{xx} a_1 \\ \Sigma_{xx}^{-1} \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma'_{xy} a_1 &= \frac{(a_1 \Sigma_{xy} b_1)^2}{(a'_1 \Sigma_{xx} a_1)(b'_1 \Sigma_{yy} b_1)} a_1 \\ \Sigma_{xx}^{-1} \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma'_{xy} a_1 &= (\rho_1^*)^2 a_1 = k_1 a_1 \end{aligned} \quad (3.23)$$

Equation 3.23 shows that a_1 is an eigenvector of $\Sigma_{xx}^{-1} \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma'_{xy}$. Thus $(\rho_1^*)^2$ is maximised when k_1 is the biggest eigenvalue of $\Sigma_{xx}^{-1} \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma'_{xy}$ and a_1 is the biggest eigenvector corresponding to the biggest eigenvalue. Also, substituting equation 3.19 into equation 3.21, b_1 is the biggest eigenvector corresponding to k_2 the biggest eigenvalue of $\Sigma_{yy}^{-1} \Sigma'_{xy} \Sigma_{xx}^{-1} \Sigma_{xy}$.

$$\begin{aligned} \Sigma'_{xy} \Sigma_{xx}^{-1} \Sigma_{xy} b_1 &= \frac{(a_1 \Sigma_{xy} b_1)^2}{(a'_1 \Sigma_{xx} a_1)(b'_1 \Sigma_{yy} b_1)} \Sigma_{yy} b_1 \\ \Sigma_{yy}^{-1} \Sigma'_{xy} \Sigma_{xx}^{-1} \Sigma_{xy} b_1 &= \frac{(a_1 \Sigma_{xy} b_1)^2}{(a'_1 \Sigma_{xx} a_1)(b'_1 \Sigma_{yy} b_1)} b_1 \\ \Sigma_{yy}^{-1} \Sigma'_{xy} \Sigma_{xx}^{-1} \Sigma_{xy} b_1 &= (\rho_1^*)^2 b_1 = k_2 b_1 \end{aligned} \quad (3.24)$$

Therefore, the first population canonical correlation coefficient is given by

$$\begin{aligned} \rho_1^* &= \sqrt{\text{the largest eigenvalue of } \Sigma_{xx}^{-1} \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma'_{xy}} \\ &= \sqrt{\text{the largest eigenvalue of } \Sigma_{yy}^{-1} \Sigma'_{xy} \Sigma_{xx}^{-1} \Sigma_{xy}}. \end{aligned} \quad (3.25)$$

The analogous sample result for equation 3.25 is

$$\begin{aligned} \hat{\rho}_1^* &= \sqrt{\text{the largest eigenvalue of } \mathbf{S}_{xx}^{-1} \mathbf{S}_{xy} \mathbf{S}_{yy}^{-1} \mathbf{S}'_{xy}} \\ &= \sqrt{\text{the largest eigenvalue of } \mathbf{S}_{yy}^{-1} \mathbf{S}'_{xy} \mathbf{S}_{xx}^{-1} \mathbf{S}_{xy}}. \end{aligned} \quad (3.26)$$

3.8.2 Deriving the second canonical variate pair

In CCA, the canonical variate pairs are selected such that each pair is highly correlated and subsequent pairs are independent of each other. The second canonical variate pair, (X_2^*, Y_2^*) , with coefficients $a_{21}, a_{22}, \dots, a_{2q}$ and $b_{21}, b_{22}, \dots, b_{2p}$ are selected such that they maximise the second canonical correlation coefficient ρ_2^* subject to the condition that the second canonical variate pair should be independent of the first canonical variate pair. This implies that the pairs (X_1^*, X_2^*) and (Y_1^*, Y_2^*) have to be uncorrelated, $cov(X_1^*, X_2^*) = cov(Y_1^*, Y_2^*) = 0$

Consider n eigenvectors e_1, e_2, \dots, e_n corresponding to n different eigenvalues. For n vectors to be linearly dependent, one vector has to be a combination of the rest as follows:

$$e_1 = \alpha_2 e_2 + \alpha_3 e_3 + \dots + \alpha_n e_n \quad (3.27)$$

such that $\alpha_2, \alpha_3, \dots, \alpha_n \neq 0$.

Applying a linear transformation to both sides of equation 3.27 we have

$$\lambda_1 e_1 = \alpha_2 \lambda_2 e_2 + \alpha_3 \lambda_3 e_3 + \dots + \alpha_n \lambda_n e_n \quad (3.28)$$

where $\lambda_1, \lambda_2, \dots, \lambda_n \neq 0$.

Dividing both sides by λ_1 we have

$$e_1 = \alpha_2 \frac{\lambda_2}{\lambda_1} e_2 + \alpha_3 \frac{\lambda_3}{\lambda_1} e_3 + \dots + \alpha_n \frac{\lambda_n}{\lambda_1} e_n. \quad (3.29)$$

Subtracting equation 3.29 from equation 3.27 we have

$$0 = \alpha_2 \left(1 - \frac{\lambda_2}{\lambda_1}\right) e_2 + \alpha_3 \left(1 - \frac{\lambda_3}{\lambda_1}\right) e_3 + \dots + \alpha_n \left(1 - \frac{\lambda_n}{\lambda_1}\right) e_n. \quad (3.30)$$

From equation 3.30 we must conclude that e_1, e_2, \dots, e_n are null vectors, which is a contradiction. We thus conclude that n eigenvectors corresponding to n distinct eigenvalues are independent. Since eigenvectors corresponding to distinct eigenvalues

are linearly independent, the eigenvector a_1 , in equation 3.23, is linearly independent to a_2 , the second largest eigenvector corresponding to the second largest eigenvalue of $\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma'_{xy}$.

Similarly, b_1 , is linearly independent to b_2 , the second largest eigenvector corresponding to the second largest eigenvalue of $\Sigma_{yy}^{-1}\Sigma'_{xy}\Sigma_{xx}^{-1}\Sigma_{xy}$.

Therefore, the second canonical correlation coefficient is given by

$$\begin{aligned}\rho_2^* &= \sqrt{\text{the second largest eigenvalue of } \Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma'_{xy}} \\ &= \sqrt{\text{the second largest eigenvalue of } \Sigma_{yy}^{-1}\Sigma'_{xy}\Sigma_{xx}^{-1}\Sigma_{xy}}.\end{aligned}\quad (3.31)$$

The analogous sample result for equation 3.31 is

$$\begin{aligned}\hat{\rho}_2^* &= \sqrt{\text{the second largest eigenvalue of } \mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-1}\mathbf{S}'_{xy}} \\ &= \sqrt{\text{the second largest eigenvalue of } \mathbf{S}_{yy}^{-1}\mathbf{S}'_{xy}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}}.\end{aligned}\quad (3.32)$$

3.8.3 Deriving the i^{th} canonical variate pair

The i^{th} canonical variate pair, (X_i^*, Y_i^*) , with coefficients $a_{i1}, a_{i2}, \dots, a_{iq}$ and $b_{i1}, b_{i2}, \dots, b_{ip}$ that maximise the i^{th} canonical correlation coefficient ρ_i^* is subject to the following constraints

$$\begin{aligned}\text{cov}(X_1^*, X_i^*) &= \text{cov}(Y_1^*, Y_i^*) = 0 \\ \text{cov}(X_2^*, X_i^*) &= \text{cov}(Y_2^*, Y_i^*) = 0 \\ &\vdots \\ \text{cov}(X_{i-1}^*, X_i^*) &= \text{cov}(Y_{i-1}^*, Y_i^*) = 0.\end{aligned}\quad (3.33)$$

Given i eigenvalues, the eigenvector a_{i-1} is linearly independent to a_i , the i^{th} largest eigenvector corresponding to the i^{th} largest eigenvalue of $\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma'_{xy}$.

Similarly, the eigenvector b_{i-1} is linearly independent to b_i , the i^{th} largest eigenvector corresponding to the i^{th} largest eigenvalue of $\Sigma_{yy}^{-1}\Sigma'_{xy}\Sigma_{xx}^{-1}\Sigma_{xy}$. Therefore, the i^{th} canonical

correlation coefficient is given by

$$\begin{aligned}\rho_1^* &= \sqrt{\text{the } i^{\text{th}} \text{ largest eigenvalue of } \Sigma_{xx}^{-1} \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma'_{xy}} \\ &= \sqrt{\text{the } i^{\text{th}} \text{ largest eigenvalue of } \Sigma_{yy}^{-1} \Sigma'_{xy} \Sigma_{xx}^{-1} \Sigma_{xy}}.\end{aligned}\quad (3.34)$$

The analogous sample result for equation 3.34 is

$$\begin{aligned}\hat{\rho}_i^* &= \sqrt{\text{the } i^{\text{th}} \text{ largest eigenvalue of } \mathbf{S}_{xx}^{-1} \mathbf{S}_{xy} \mathbf{S}_{yy}^{-1} \mathbf{S}'_{xy}} \\ &= \sqrt{\text{the } i^{\text{th}} \text{ largest eigenvalue of } \mathbf{S}_{yy}^{-1} \mathbf{S}'_{xy} \mathbf{S}_{xx}^{-1} \mathbf{S}_{xy}}.\end{aligned}\quad (3.35)$$

3.9 Canonical weights

The *canonical weights or raw correlation coefficients* measure the amount of contribution each variable makes to a variate. Raw correlation coefficients are sensitive to scaling and are thus not appropriate for interpretation. For the first canonical variate pair, (X_1^*, Y_1^*) , the raw correlation coefficients $a_{11}, a_{12}, \dots, a_{1q}$ and $b_{11}, b_{12}, \dots, b_{1p}$ are selected such that they maximise the first canonical correlation coefficient ρ_1^* .

3.10 Standardized coefficients

Multiplying a_i and b_i by the standard deviations of corresponding variables removes the effect of scaling.

$$c_i = D_x a_i, \quad d_i = D_y b_i \quad (3.36)$$

where $D_x = \text{diag}(\Sigma_{x1}, \Sigma_{x2}, \dots, \Sigma_{xq})$ and $D_y = \text{diag}(\Sigma_{y1}, \Sigma_{y2}, \dots, \Sigma_{yp})$.

The eigenvectors of the matrices $R_{xx}^{-1} R_{xy} R_{yy}^{-1} R_{xy}$ and $R_{yy}^{-1} R_{xy} R_{xx}^{-1} R_{xy}$ give the sample estimates \hat{c}_i and \hat{d}_i respectively.

The coefficients in c_i namely $c_{i1}, c_{i2}, \dots, c_{iq}$ depict the amount of contribution made by

each of X_1, X_2, \dots, X_q to X_i^* and the coefficients in d_i namely $d_{i1}, d_{i2}, \dots, d_{ip}$ represent the amount of variation contributed by each of Y_1, Y_2, \dots, Y_q to Y_i^* (Alvin, 2002).

3.11 Canonical loadings

Canonical loadings are the correlations between the variables and variates within the same group. CCA generates multiple dimension of relationships between variates. Each relationship is independent of the others (Kabir et al., 2014). The canonical loadings fluctuate from dimension to dimension representing a variable's contribution to the given relationship.

The loadings for the X - set are given by

$$R_{xx}\hat{c}_i \tag{3.37}$$

and the loadings for Y - set are given by

$$R_{yy}\hat{d}_i. \tag{3.38}$$

3.12 Canonical cross loadings

Canonical cross loadings are correlations between the variables and variates within the different groups. In other words, it is the correlation between the independent variables and the dependent variate or the correlation between the dependent variables and the independent variate. This measure is obtained by multiplying canonical loadings with canonical correlation coefficients.

The cross loadings for X - set are given by

$$R_{xx}\hat{c}_i\hat{\rho}_i^* \tag{3.39}$$

and the cross loadings for Y - set are given by

$$R_{yy}\hat{d}_i\hat{\rho}_i^*. \quad (3.40)$$

They offer more direct interpretations than conventional loadings.

3.13 Canonical variate scores

The canonical variate scores of X - set and Y - set of variables from the i^{th} canonical variate pair (X_i^*, Y_i^*) are, respectively, Xc_i and Yd_i where X and Y are vectors of predictors and criterion variables.

The scores of X_i^* can be used to predict Y_i^* . This predicted value is obtained from the regression analysis of Y_i^* on X_i^* . The predicted Y_i^* is given by

$$\hat{Y}^*i = \rho_i(X_i^* - \hat{c}'_i\bar{X}_i) + \hat{d}'_i\bar{Y} \quad (3.41)$$

The predicted value of Y_i^* is obtained from the regression analysis of Y_i^* on Y_i^* (Bhuyan, 2005).

3.14 Tests of significance

3.14.1 Tests of independence between X - set and Y - set

In order to perform CCA, the very first thing to determine is if two groups of variables are dependent. We wish to test the null hypothesis that the canonical coefficients corresponding to each variable are all equal to zero. This is comparable to the null hypothesis that the X - set is independent of the Y - set. (Alvin, 2002). The test statistic is Wilk's lambda Λ .

Wilks' lambda Λ is the ratio

$$\Lambda(p, n - 1 - q, q) = \prod_{i=1}^k (1 - l_i) = \frac{|S_{yy}^{-1} S_{yx} S_{xx}^{-1} S_{xy}|}{|S_{yy}|} \quad (3.42)$$

where $k = \min(p, q)$ and l_i is the i^{th} eigenvalue of $S_{yy}^{-1} S_{yx} S_{xx}^{-1} S_{xy}$. If the values of these statistics are too large, the p-value is small. This indicates rejection of the null hypothesis

$$H_o : \Sigma_{xy} = \mathbf{O}$$

and can conclude that the X – set and the Y – set are dependent. Also, the above null hypothesis is comparable to testing the null hypothesis that all variate pairs are not correlated,

$$H_o : \rho_1^* = \rho_2^* = \dots = \rho_p^*.$$

For a large n, the statistic Λ in equation 3.42 follows a Chi-square distribution with pq degrees of freedom, where,

$$\chi^2 = -[(n - 1) - \frac{1}{2}(p + q + 1)] \ln \Lambda. \quad (3.43)$$

We reject H_o if $\chi^2 \geq \chi_\alpha^2$ and hence perform CCA.

Alternatively, the F -approximate of equation 3.42 given by

$$F = \frac{1 - \Lambda^{\frac{1}{t}}}{\Lambda^{\frac{1}{t}}} \frac{df_2}{df_1}, \quad (3.44)$$

which follows a F -distribution (Alvin, 2002). The degrees of freedom are df_1 and df_2 , where $df_1 = pq$, $df_2 = wt - \frac{1}{2}pq + 1$ and

$$t = \sqrt{\frac{p^2 q^2 - 4}{p^2 + q^2 - 5}}.$$

We reject H_o if $F > F_\alpha$. There are other test statistics for the hypothesis $H_o : \Sigma_{xy} = \mathbf{O}$ listed below;

1. Pillai's test statistic of independence

$$V^{(s)} = \sum_{i=1}^s l_i \quad (3.45)$$

where $s = \min(p, q)$ (Alvin, 2002).

The approximate F - statistic is

$$F = \frac{(2N + s + 1)V^{(s)}}{(2m + s + 1)(s - V^{(s)})} \quad (3.46)$$

which is approximated as $F_{s(2m+s+1), s(2N+s+1)}$ where $m = \frac{1}{2}(|q - p| - 1)$ and $N = \frac{1}{2}(n - q - p - 2)$.

2. The Lawley-Hotelling statistic of independence

$$U^{(s)} = \sum_{i=1}^s \frac{l_i}{1 - l_i}. \quad (3.47)$$

The approximate F - statistic is

$$F = \frac{2(sN + 1)U^{(s)}}{s^2(2m + s + 1)} \quad (3.48)$$

with $s(2m + s + 1)$ and $2(sN + 1)$ degrees of freedom.

3. Roy's largest root

$$\theta = l_i. \quad (3.49)$$

3.14.2 Tests of significance of the i^{th} variate

If the test in equation 3.42 rejects H_o , certainty of significance of the remaining canonical correlations is not guaranteed (Alvin, 2002). There are $k = \min(p, q)$ canonical variate pairs and the significance of these canonical variate pairs is judged by Wilk's lambda Λ as well.

For the first canonical pair, Wilk's lambda is given by

$$\Lambda_1 = \prod_{i=1}^s (1 - l_i). \quad (3.50)$$

To obtain the Wilk's lambda for the second canonical variate pair, we delete the influence of the first canonical variate pair for Λ to obtain

$$\Lambda_2 = \prod_{i=2}^s (1 - l_i). \quad (3.51)$$

If the null hypothesis in equation 3.52 is rejected, we conclude that as a minimum ρ_2^* is significantly not equal to zero. We proceed in this style, testing each ρ_i^* one at a time, until a test fails to reject the null hypothesis (Alvin, 2002). The test statistic used when the test is done k times is given by

$$\Lambda_k = \prod_{i=k}^s (1 - l_i). \quad (3.52)$$

3.15 Assumptions of Canonical Correlation Analysis

CCA does not require strict adherence to some assumptions. However, if assumptions are taken into consideration, the interpretation of relationships is enhanced. The following are the important assumptions of CCA.

- (i) Multiple continuous or categorical variables for both dependent variables and independent variables must be available from the data in order to perform CCA
- (ii) The two sets of variables must have a linear relationship
- (iii) Each variable from the two sets should be normally distributed
- (iv) The relationships between groups of variables should be homoscedastic
- (v) There should be no multicollinearity among independent variables

CHAPTER 4

DATA ANALYSIS, RESULTS AND DISCUSSIONS

4.1 Introduction

This chapter contains the exploratory data analysis (EDA) results from the data used in the study. EDA checks the assumptions of canonical correlation analysis and also provide descriptive information about the data. The results from the computations of correlations, eigenvalues, canonical weights, canonical loading and canonical cross loadings have also been presented. A discussion and interpretation of the results has also been included in this chapter.

4.2 Exploratory Data Analysis

The summary statistics of the data set are shown in Table 4.1.

TABLE 4.1: Summary statistics for KCSE examination results in STEM subjects and school characteristics data

Variables	N	Mean	Median	Minimum	Maximum	Std. Deviation
<i>Predictor set</i>						
X_1	77	0.24	0.22	0.10	0.73	0.109
X_2	77	127.71	115.00	22.00	339.00	74.923
X_3	77	0.86	0.87	0.52	1.00	0.146
X_4	77	0.33	0.27	0.04	1.00	0.213
X_5	77	0.36	0.21	0.00	1.00	0.357
<i>Criterion Set</i>						
Y_1	77	4.30	3.24	1.27	10.73	2.783
Y_2	77	5.10	4.40	1.82	11.21	2.346
Y_3	77	5.10	4.57	1.40	10.30	2.332
Y_4	77	4.86	4.14	1.55	11.06	2.437

The predictor set (X - set) of variables represents the school characteristics and includes

the following variables:

X_1 - Teacher to student ratio

X_2 - School size

X_3 - Proportion of students taking biology

X_4 - Proportion of students taking physics

X_5 - Proportion of students with mean scores above C+.

The criterion set (Y - set) of variables represents the level of performance in career subjects in STEM education and includes the following variables:

Y_1 - KCSE Mathematics mean score

Y_2 - KCSE Biology mean score

Y_3 - KCSE Physics mean score

Y_4 - KCSE Chemistry mean score.

X_1 is a measure of staffing in schools and is the ratio of number of teachers to the number of candidates. X_1 ranges from 0.10 to 0.73. X_2 is a measure of school size and is the number of candidates in a school. Mean scores are values ranging from 1 to 12. The lowest mean was obtained in mathematics while the highest mean was obtained in biology. From the sample, all the students participated in mathematics and chemistry. 86 percent of the students participated in biology and 33 percent participated in physics. The highest mean score obtained amongst the schools was 11.21 in biology and the lowest was 1.27 in mathematics.

The Figure 4.1 shows the scatter plot for the criterion set of variables. It depicts linearity in the relationship between variables in the Y-set. CCA measures a linear relationship between the variables. The criterion variables are correlated, making CCA appropriate for this data.

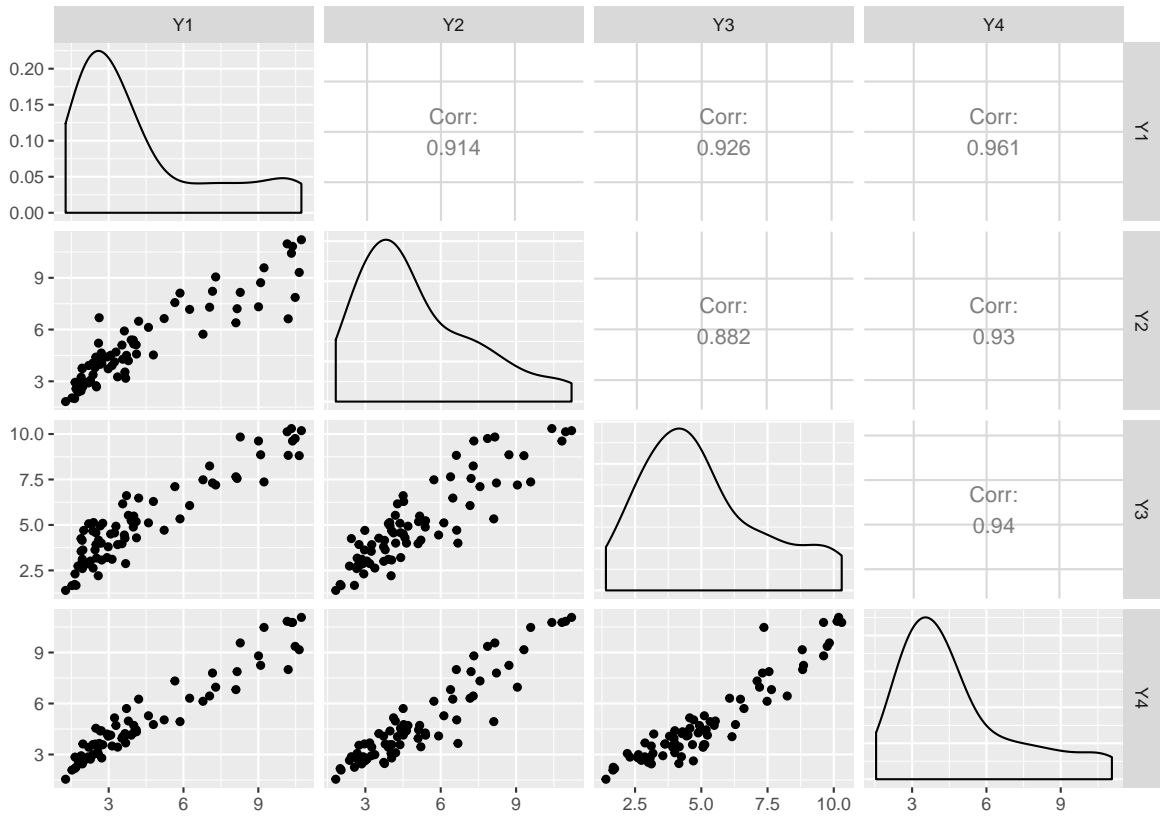


FIGURE 4.1: Scatter plot for the criterion set of variables

The Figure 4.2 shows the scatter plot for the predictor set of variables. It shows that the predictor variables are highly skewed. CCA can still accommodate non-normal variables if the distribution form does not decrease the correlation with other variables. The scatter plot depicts that none of the variables can be predicted by the rest of the variables since the distributions do not follow any systematic pattern.

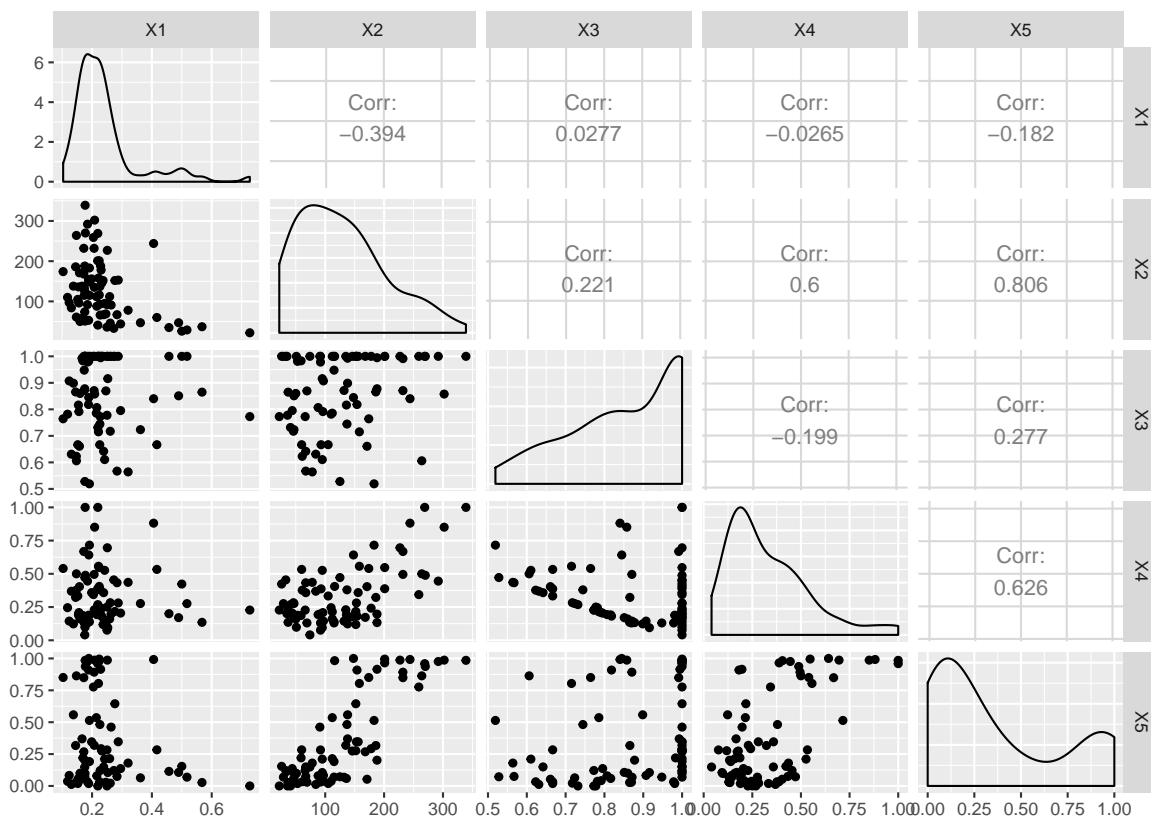


FIGURE 4.2: Scatter plot for the predictor set of variables

4.3 Canonical Correlation Analysis

4.3.1 Correlations

The correlations between all the variables in the study are shown in table 4.2. Most correlations are significant at the 0.01 level. The teacher to student ratio (X_1) correlates negatively(-0.394) with the school size (X_2). This indicates that increases in the number of students are not proportional to the increases in the number of teachers in public schools. The teacher to student ratio reduces with larger school sizes. The school size (X_2) correlates highly with all variables except the proportion of students taking biology (X_3). This indicates that the level of performance in biology is not significantly affected by the size of the school. The proportion of students taking physics (X_4) is highly correlated with the mean score in chemistry (Y_4). This shows that higher proportions of students taking physics correspond to better scores in chemistry. The

TABLE 4.2: Correlations within and between the predictor and criterion two sets of variables

	X_1	X_2	X_3	X_4	X_5	Y_1	Y_2	Y_3	Y_4
X_1	1								
X_2	-.394**	1							
X_3	0.028	0.221	1						
X_4	-0.026	.600**	-0.199	1					
X_5	-0.182	.806**	.277*	.626**	1				
Y_1	-0.156	.813**	.283*	.688**	.958**	1			
Y_2	-0.159	.798**	.289*	.664**	.918**	.914**	1		
Y_3	-0.217	.790**	.459**	.544**	.919**	.926**	.882**	1	
Y_4	-0.144	.798**	.333**	.700**	.938**	.961**	.930**	.940**	1

** . Correlation is significant at the 0.01 level (2-tailed).

* . Correlation is significant at the 0.05 level (2-tailed).

proportion of students with mean scores above C+ (X_5) is very highly correlated with the mathematics mean score (Y_1). The mathematics mean score (Y_1) and biology mean score (Y_2) correlate very highly with the mean score in chemistry (Y_4). The physics mean score (Y_3) correlate very highly with the proportion of students taking biology (X_3) and the chemistry mean score (Y_4).

4.3.2 Test of independence between X - set and Y - set

To test overall model fit, the null hypothesis that the X - set and Y - set are independent is tested. The results of tests of multivariate significance of canonical correlation are displayed in Table 4.3. Pillai's, Helling's, Wilk's and Roy's tests all confirm that at least one variate pair is significant with $p < .05$.

TABLE 4.3: Multivariate Tests of Significance (S = 4, M = 0, N = 33)

Test Name	Value	Approx. F	Hypoth. DF	Error DF	Sig. of F
Pillais	1.46726	8.22629	20	284	0.000
Hotellings	23.63956	78.60153	20	266	0.000
Wilks	0.0227	24.13038	20	226.48	0.000
Roys	0.95814				

4.3.3 Test of significance of the second variate pair

Wilks' lambda Λ is used to test the significance of each of $k = \min(p, q)$ variates. In this study, the number of variables in the X - set was $q = 5$ and the number of variables in the Y - set was $p = 4$. Thus, the number of possible variates is 4. From the results in Table 4.4 only the first two variates are significant at $p < .05$. CCA thus provides two canonical roots or dimensions that help describe the linear relationships between school characteristics and STEM education.

TABLE 4.4: Dimension Reduction Analysis

Variates	Wilks L.	F	Hypoth. DF	Error DF	Sig. of F
1 TO 4	0.023	24.13038	20	226.48	0.000
2 TO 4	0.542	3.96469	12	182.85	0.000
3 TO 4	0.864	1.7704	6	140	0.109
4 TO 4	0.994	0.2177	2	71	0.805

4.3.4 Eigenvalues and Canonical Correlations

The canonical correlation coefficients and the eigenvalues of the canonical roots are reported in Table 4.5. Canonical correlation coefficients are the square roots of eigenvalues.

TABLE 4.5: Eigenvalues and Canonical Correlations

Variates	Eigenvalue	Canon Cor.
1	0.95814	0.97885
2	0.37225	0.61012
3	0.13078	0.36163
4	0.0061	0.07807

The first eigenvalue is 0.95814 and its corresponding canonical correlation coefficient estimate is $\hat{\rho}_1^* = 0.97885$. The correlation between the first variate pair is highly significantly correlated. This means that at least one variable in the X - set correlates significantly with at least one variable in the Y - set.

The second eigenvalue is 0.37225 and its corresponding canonical correlation coefficient

estimate is $\hat{\rho}_2^* = 0.61012$. This means that another variable in the X - set correlates significantly with another variable in the Y - set.

The third and fourth variates are not significant at $p < .05$ and thus their corresponding canonical correlation coefficients and eigenvalues are not interpreted.

4.3.5 Canonical weights

The raw canonical weights (or coefficients) are interpreted like coefficients in linear regression. However, since the variables in this study have different sizes, we interpret the standardized canonical coefficients. Standardized canonical coefficients do not reflect the differences in scaling and are hence used in the canonical function to calculate the canonical variate scores. The raw and standardized canonical weights are displayed in Table 4.6.

TABLE 4.6: Raw and Standardized canonical weights

Variables	Variate 1		Variate 2	
	Raw canonical weights	Standardized canonical weights	Raw canonical weights	Standardized canonical weights
Predictor set				
X1	-0.03319	-0.00361	-3.12329	-0.34021
X2	0.00102	0.07638	0.00083	0.06201
X3	1.03985	0.15201	4.12889	0.6036
X4	0.99701	0.21264	-3.01848	-0.64379
X5	2.09484	0.74749	0.38857	0.13865
Criterion set				
Y1	0.17336	0.48242	-0.44137	-1.22826
Y2	0.08885	0.20844	-0.01315	-0.03086
Y3	0.06253	0.14582	1.28775	3.00309
Y4	0.07734	0.18845	-0.67646	-1.64821

Based on the data, the first canonical variate pair is

$$\hat{X}_1^* = \hat{c}'_1 X = -0.00361X_1 + 0.07638X_2 + 0.15201X_3 + 0.21264X_4 + 0.74749X_5$$

$$\hat{Y}_1^* = \hat{d}'_1 Y = 0.48242Y_1 + 0.20844Y_2 + 0.14582Y_3 + 0.18845Y_4. \quad (4.1)$$

For the first variate pair, the the proportion of students with C+ and above (X_5) and the mean score for mathematics Y_1 contribute the most to canonical correlation.

The variate scores of X - set and Y - set of variables from the first canonical variate pair are obtained by substituting the variable observations from the sample into equation 4.1. The second canonical variate pair is

$$\hat{X}_2^* = \hat{c}'_2 X = -0.34021X_1 + 0.06201X_2 + 0.6036X_3 - 0.64379X_4 + 0.13865X_5$$

$$\hat{Y}_2^* = \hat{d}'_2 Y = -1.22826Y_1 - 0.03086Y_2 + 3.00309Y_3 - 1.64821Y_4. \quad (4.2)$$

For the second variate pair, the the proportion of students taking physics (X_4) and the mean score for physics (Y_3) contribute the most to canonical correlation.

The variate scores of X - set and Y - set of variables from the first canonical variate pair are obtained by substituting the variable observations from the sample into equation 4.2.

4.3.6 Canonical loadings

The canonical loadings are correlations between variable scores and variables in the same domain. In Table 4.7 the canonical variate loadings for this study are presented. Although canonical loadings may appear to demonstrate some similarity with canonical weights, there are important differences due to multicollinearity. For the first variate, the all canonical loadings of the X - set exceed 0.3 apart from the loading for the teacher to student ratio (X_1). The rest of the school factors correlate positively. The variable with the largest loading is the proportion of students with C+ and above (X_5). The canonical loadings of the Y - set all exceed 0.3 and are positive. This shows that the measures of performance in STEM education are highly positively correlated. The variable with the largest loading is the mean score for mathematics Y_1 .

For the second variate, the X - set variables with the largest loadings are the proportion

TABLE 4.7: Canonical Loadings

Variables	Variate	
	1	2
Predictor set		
X1	-0.17084	-0.35606
X2	0.84178	0.05465
X3	0.3337	0.7743
X4	0.69637	-0.63077
X5	0.98501	0.01467
Criterion set		
Y1	0.98894	-0.05801
Y2	0.95308	-0.03648
Y3	0.95363	0.28957
Y4	0.98274	-0.03521

of students taking biology (X_3) and the proportion of students taking physics (X_4). The proportion of students taking biology is positively correlated with the school factors. However, the proportion of students taking physics is negatively correlated with the school factors. The variables of the Y - set all have loadings less than 0.3, with the largest loading being the mean score for physics Y_3 .

These results are similar to the ones obtained from canonical weights. We hence conclude that multicollinearity does not confound the ability of CCA to isolate the most influential variable from the sample data.

The interpretation of the canonical loadings from the first variate pair is that the proportions of students scoring grade C+ and above is the most influential variable in defining school characteristics. Also, the mean score in mathematics is the most influential variable in defining the level of performance in STEM education. From the second variate pair, the proportions of students taking biology and physics are the second most influential variables in defining school characteristics. Additionally, the mean score in physics is the second most influential variable in defining the level of performance in STEM education.

TABLE 4.8: Canonical Cross Loadings

Variables	Variate	
	1	2
Predictor set		
X1	-0.16722	-0.21724
X2	0.82397	0.03334
X3	0.32664	0.47241
X4	0.68164	-0.38484
X5	0.96418	0.00895
Criterion set		
Y1	0.96802	-0.03539
Y2	0.93292	-0.02225
Y3	0.93346	0.17667
Y4	0.96196	-0.02148

4.3.7 Canonical cross loadings

The measures of the relationship between any variable in the Y - set and any variable in the X - set appear in Table 4.8.

For the first variate pair, it is seen that the proportion of students with C+ and above (X_5) is the highest correlated variable with the variables in Y - set. This implies that the level of academic performance in STEM education is mostly influenced by the proportion of students with C+ and above based on the data used in this study. For the second variate pair, it is observed that the proportions of students taking biology (X_3) and those taking physics (X_4) are the highest correlated variable with the variables in Y - set. This implies that the level of academic performance in STEM education is mostly influenced by the proportion of students taking biology and physics based on the data used in this study.

4.3.8 Prediction

The two sets of variate scores obtained in equation 4.1 can be used to study the relationship between school characteristics and performance in STEM education i.e. the variables in X - set and Y - set. The score X_1^* can be used to predict the score Y_1^* ,

where the predicted score is given by

$$\hat{Y}_1^* = \sqrt{l_1}(X_1^* - \hat{c}'_1 \hat{X}) + \hat{d}'_1 \hat{Y} \quad (4.3)$$

Additionally, the predicted score of Y_2^* is

$$\hat{Y}_2^* = \sqrt{l_2}(X_2^* - \hat{c}'_2 \hat{X}) + \hat{d}'_2 \hat{Y}. \quad (4.4)$$

For the present data, we have

$$\begin{aligned} l_1 &= 0.95814, \quad l_2 = 0.37225, \\ \hat{X} &= \begin{pmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \bar{X}_3 \\ \bar{X}_4 \\ \bar{X}_5 \end{pmatrix} = \begin{pmatrix} 0.2380 \\ 127.71 \\ 0.8614 \\ 0.3320 \\ 0.3605 \end{pmatrix}, \quad \hat{Y} = \begin{pmatrix} \bar{Y}_1 \\ \bar{Y}_2 \\ \bar{Y}_3 \\ \bar{Y}_4 \end{pmatrix} = \begin{pmatrix} 4.3019 \\ 5.0953 \\ 5.1027 \\ 4.8628 \end{pmatrix}, \\ \hat{c}'_1 &= \begin{pmatrix} -0.00361, 0.07638, 0.15201, 0.21264, 0.74749 \end{pmatrix}, \\ \hat{c}'_2 &= \begin{pmatrix} -0.34021, 0.06201, 0.6036, -0.64379, 0.13865 \end{pmatrix}, \\ \hat{d}'_1 &= \begin{pmatrix} 0.48242, 0.20844, 0.14582, 0.18845 \end{pmatrix}, \\ \hat{d}'_2 &= \begin{pmatrix} -1.22826, -0.03086, 3.00309, -1.64821 \end{pmatrix}. \end{aligned}$$

Hence,

$$\begin{aligned} \hat{Y}_1^* &= 0.9788[-0.00361X_1 + 0.07638X_2 + 0.15201X_3 + 0.21264X_4 + 0.74749X_5 - 10.225] + 4.798 \\ &= -0.00353X_1 + 0.07476X_2 + 0.14879X_3 + 0.20813X_4 + 0.73164X_5 - 5.21 \quad (4.5) \end{aligned}$$

and

$$\hat{Y}_2^* = 0.6101[-0.34021X_1 + 0.06201X_2 + 0.6036X_3 - 0.64379X_4 + 0.13865X_5 - 8.195] + 1.868$$

$$= -0.20756X_1 + 0.03783X_2 + 0.36826X_3 - 0.39278X_4 + 0.08459X_5 - 3.1318 \quad (4.6)$$

From equation 4.5 the variable that contributes the most is the proportion of students with mean scores above C+. The mean scores of mathematics, biology, physics and chemistry would be predicted by obtaining the eigenvector corresponding to the first eigenvalue of $\mathbf{S}_{yy}^{-1}\mathbf{S}'_{xy}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$.

From equation 4.6 the variables that contributes the most are the proportions of students taking biology and physics. The mean scores of mathematics, biology, physics and chemistry would be predicted by obtaining the eigenvector corresponding to the second eigenvalue of $\mathbf{S}_{yy}^{-1}\mathbf{S}'_{xy}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$.

CHAPTER 5

SUMMARY OF FINDINGS, CONCLUSIONS AND RECOMMENDATIONS

5.1 Summary of Findings

Data from 9,834 candidates from 77 public secondary schools in Nairobi County revealed that school factors significantly correlate with the level of performance in STEM education. Canonical correlation analysis extracted two significant canonical variate pairs with canonical correlations 0.97885 and 0.61012 respectively. The independent variables with the largest canonical loadings were the proportion of students with C+ and above in the first variate and the proportions of students taking biology and physics in the second variate. The dependent variables with the largest canonical loadings were the mean score for mathematics in the first variate and the mean score for physics in the second variate. For the first variate pair, it is seen that the proportion of students with C+ and above is the highest correlated variable with the variables in Y - set. For the second variate pair, the proportions of students taking biology and those taking physics are the highest correlated variables with the variables in Y - set.

5.2 Conclusions

Based on the standardised canonical coefficients and the canonical loadings the most contributing subject in defining the level of performance in STEM education is mathematics. Despite the fact that physics is optional and has the lowest mean proportion of participation, it is the second most contributing subject in defining STEM education. Similarly, the most influential variable in defining school characteristics with regard to STEM education is the school's proportion of students with C+ and above. Schools that have larger proportions of students with C+ and above perform better in mathematics.

The proportions of students taking biology and physics are the other highly influential variables in defining school characteristics that support STEM education. Physics is performed better when fewer students opt to study it, implying that smaller class sizes are most ideal for better scores in physics.

5.3 Recommendations

To improve the level of performance in STEM career subjects, administrators of schools should strive to increase the proportions of students scoring C+ and above. Interventions should be sought in order to facilitate the provision of adequate staffing in physics so as to improve the participation and performance in physics.

Further studies should be done to establish the relationship between individual factors and participation and performance in STEM career subjects.

REFERENCES

- Albert, Owino Ogutu, Ahmed Osman, and Alice Yungungu (2014). “An investigation of Factors that Influence Performance in KCSE Biology in selected secondary schools in Nyakach District , Kisumu County , Kenya The problem of poor performance in science subjects is global as indicated by studies done by Valverde and Sch”. In: 3.2, pp. 957–977.
- Alvin, C Rencher (2002). “Methods of multivariate analysis”. In: *Wiley Interscience*.
- Bhuyan, KC (2005). *Multivariate Analysis & Its Applications*. New Central Book Agency.
- Bickel, Robert (2007). *Multilevel analysis for applied research: it's just regression!* Guilford Press.
- Bryant, Mykeko (2012). *Cracking the code*. Vol. 42. 8, pp. 16–17. ISBN: 9789231002335.
- Carnevale, Anthony P et al. (2011). “Career Clusters: Forecasting Demand for High School through College Jobs, 2008-2018.” In: *Georgetown University Center on Education and the Workforce*.
- Gonzalez, Heather B and Jeffrey J Kuenzi (2012). “Science, technology, engineering, and mathematics (STEM) education: A primer”. In: Congressional Research Service, Library of Congress.
- Huyer, Sophia (2015). “Is the gender gap narrowing in science and engineering”. In: *UNESCO science report: towards 2030*, p. 85.
- Kabir, Alamgir et al. (2014). “Canonical correlation analysis of infant’s size at birth and maternal factors: a study in rural Northwest Bangladesh”. In: *PloS one* 9.4, e94243.
- Mbaki, Lydia, Musau Joash, and James Matee Muola (2010). “Determinants of girls ’ performance in science , mathematics and technology subjects in public secondary schools in Kenya”. In: *Int. J. Educ. Admin. Pol. Stud.* 5.3, pp. 33–42.
- Nagel, Sarah (2017). “Policy Analysis on UNESCO’s Action Plan for Gender Equality 2014-2021”. In:

- Patton, Wendy and Mary McMahon (2014). *Career development and systems theory: Connecting theory and practice*. Vol. 2. Springer.
- Solon, Gary, Steven J Haider, and Jeffrey M Wooldridge (2015). “What are we weighting for?” In: *Journal of Human resources* 50.2, pp. 301–316.
- The World Bank (2016). *World Development Report 2016: Digital Dividends*. Vol. 65. 3, pp. 461–468. ISBN: 978-1-4648-0671-1.
- Thompson, Bruce (1991). “A primer on the logic and use of canonical correlation analysis.” In: *Measurement and Evaluation in Counseling and Development*.
- Win, Rosemary and Paul W. Miller (2005). “The Effects of Individual and School Factors on University Students’ Academic Performance”. In: *The Australian Economic Review* 38.1, pp. 1–18. ISSN: 0004-9018.

APPENDICES

APPENDIX A

THE SIXTEEN CAREER CLUSTERS AND THEIR CORRESPONDING CAREER PATHWAYS

TABLE A.1: The sixteen career clusters and their corresponding career pathways

Career Cluster	Career pathways
1 Agriculture, Food and Natural Resources	Agribusiness Systems Animal Systems Environmental Service Systems Food Products and Processing Systems Natural Resources Systems Plant Systems Power, Structural and Technical Systems
2 Hospitality and Tourism	Lodging Recreation, Amusements and Attractions Restaurants and Food/ Beverage Services Travel and Tourism
3 Architecture and Construction	Construction Design/ Pre-Construction Maintenance/ Operations
4 Human Services	Consumer Services Counseling and Mental Health Services Early Childhood Development and Services Family and Community Services Personal Care Services
5 Arts, A/V Technology, and Communications	A/V Technology and Film Journalism and Broadcasting Performing Arts Printing Technology Telecommunications Visual Arts
6 Information Technology	Information Support and Services Network Systems Programming and Software Development Web and Digital Communications
7 Business Management and Administration	Administrative Support Business Information Management General Management Human Resources Management Operations Management
8 Law, Public Safety, Corrections and Security	Correction Services Emergency and Fire Management Services Law Enforcement Services Legal Services Security and Protective Services
9 Education and Training	Administration and Administrative Support

Table A.1 continued from previous page

Career Cluster	Career pathways
	Professional Support Services
	Teaching/ Training
10 Manufacturing	Healthy, Safety and Environmental Assurance Logistics and Inventory Control Maintenance, Installation and Repair Manufacturing Production Process Development Production Quality Assurance
11 Finance	Accounting Banking Services Business Finance Insurance Securities and Investments
12 Marketing	Marketing Communications Marketing Management Marketing Research Merchandising Professional Sales
13 Government and Public Administration	Foreign Service Governance National Security Planning Public Management and Administration Regulation Revenue and Taxation
14 Health Science	Biotechnology Research and Development Diagnostic Services Healthy Information Support Services Therapeutic Services
15 Science, Technology, Engineering and Mathematics	Engineering and Technology Science and Mathematics
16 Transportation, Distribution and Logistics	Facility and Mobile Equipment Maintenance Health, Safety and Environmental Management Logistics Planning and Management Services Sales and Service Transportation Operations Transportation Systems Warehousing and Distribution Center Operations

APPENDIX B

KCSE NATIONAL PERFORMANCE PER SUBJECT BETWEEN 2010 AND 2015

TABLE B.1: KCSE national performance per subject between 2010 and 2015

Subject	Overall mean					
	2010	2011	2012	2013	2014	2015
English	39.26	36.74	38.13	35.23	47.68	40.29
Kiswahili	44.34	49.01	36.32	39.91	47.68	47.88
Mathematics	19.17	21	25.3	25.1	24.02	26.88
Biology	26.71	31.72	25.38	28.7	31.83	34.8
Physics	31.5	32.94	32.53	36.87	38.84	43.68
Chemistry	22.89	23.4	27.72	25.45	32.16	34.36
History	41.73	38.45	37.14	41.78	53.83	51.71
Geography	33.86	38.15	43.09	41.02	44.02	43.92
CRE	46.05	49.38	44.34	51.93	53.15	52.48
Agriculture	31.25	34.26	32.03	31.94	40.82	44.81
Business	37.28	42.61	51	53.64	46.82	43.76