International Journal of Computational and Theoretical Statistics
ISSN (2210-1519)
Int. J. Comp. Theo. Stat. 5, No. 2 (Nov-2018)

http://dx.doi.org/10.12785/ijcts/050201

Modelling School Factors and Performance in Mathematics and Science in Kenyan Secondary Schools Using Canonical Correlation Analysis

Jeremiah Mbaria Mucunu[1]and George Muhua[2]

[1,2]School of Mathematics, University of Nairobi, Nairobi County, Kenya

Received January 17, 2018, Revised May 31, 2018, Accepted June 28, 2018, Published November 1, 2018

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.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 these subjects, this study endeavors to establish the relationship between the school characteristics and academic achievement in STEM education.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. The research utilized data from 9,834 candidates of year 2015 Kenya Certificate of Secondary Education (KCSE) from 77 public secondary schools in Nairobi County. Canonical Correlation Analysis (CCA) is a multivariate data analysis technique that seeks to establish whether two sets of variables, 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 standardized 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 contribute very highly to the level of performance in STEM education.

Keywords: Canonical correlation analysis, School factors, STEM

1. BACKGROUND INFORMATION

1.1 Introduction

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. Career development takes place in five phases based on the child's age [12]. 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 [6]. 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. 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 [7]. 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 [8].

E-mail address: mbaria2704@gmail.com, odweso@uonbi.ac.ke

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. 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 [14].

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 directs 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) emphasizes the importance of ensuring that curriculum is sensitive to gender with regards to STEM education so as to achieve Kenya Vision 2030 [11]. 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 [16]. 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 [5].

1.2 Objectives of the study

The main objective of the study was to establish the relationship between Kenyan secondary schools' characteristics and students' level of academic performance in science and mathematics. The specific objectives of the study were:

a) To determine the most contributing subject in defining the level of performance in science and mathematics

b) To determine the most influential variable in defining school characteristics with regard to STEM education

c) To determine the magnitude of the relationship that exists between school characteristics and performance in science and mathematics

d) To predict the performance in science and mathematics given the school characteristics

2. Literature Review

Albert, Ahmed and Alice [1] 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, Musau and James [10] 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. A series of correlation analyses were performed 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 Paul [16] 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, the proportion of graduates from that school and the percentage of students who passed the entry examinations from that school. A linear regression model was used to analyze 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'.

Albert, Ahmed and Alice [1] focus 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, Musau and James [10] 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, Musau and James [10] 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 not taken by all students in the given school, 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 [13].

Win and Paul [16] 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 [4].

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. When several tests in statistics are applied for each dependent variable, the probability of making a Type I error increases [15]. CCA reduces the probability of having Type I errors.

3. METHODOLOGY

The data used in this research were 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. 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.

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 [15]. CCA does not require strict adherence to some assumptions. However, if assumptions are taken into consideration, the interpretation of relationships is enhanced. The important assumptions of CCA are as follows: multiple continuous or categorical variables for both dependent variables and independent variables must be available from the data in order to perform CCA, the two sets of variables must have a linear relationship, each variable from the two sets should be normally distributed, the relationships between groups of variables should be homoscedastic and there should be no multicollinearity among independent variables.

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 linear combinations $X^* = a^{'}X$ of the variables in the X-set and $Y^* = a^{'}Y$ of the variables in the Y-set, where a and b are two vectors of constants of elements q and p respectively, are referred to as **canonical variates**. 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**. 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 [3]. 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 ith canonical variate pair is defined by (X_i, Y_i). Thus, the first canonical variate pair is given by (X_1, Y1). The canonical correlation coefficient for the ith pair of variates (1) is the correlation between X_i^* and Y_i^* where S_{xy}, S_{xx} and S_{yy} are covariance matrices

$$\rho_i^* = \frac{\mathrm{cov}(X_i^*, Y_i^*)}{\sqrt{\mathrm{var}(X_i^*)\,\mathrm{var}(Y_i^*)}} = \frac{\hat{a}_i' S_{xy} \hat{b}_i}{\sqrt{\hat{a}_i' S_{xx} \hat{a}_i}\sqrt{\hat{b}_i' S_{yy} \hat{b}_i}} \tag{1}$$

where ρ_i^* is the vector of eigenvalues and a_i and b_i are eigenvectors corresponding to the eigenvalues. CCA formulates an equation linking the X and Y variables that maximizes the canonical correlation coefficient between the pair of variates. Equating the partial derivatives of the square of equation (1) to zero, we obtain the values of a_i and b_i.

3.1 Canonical weights, loadings and cross loadings

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, the raw correlation coefficients a11, a12..., a1q and b11, b12..., a1p are selected such that they maximize the first canonical correlation coefficient ρ_1^*. Standardized correlation coefficients remove the effect of scaling and are obtained from multiplying a_i and b_i by the standard deviations of corresponding variables [2].

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 [9]. 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 $R_{xx}\hat{c}_i$ and the loadings for Y - set are given by $R_{yy}\hat{d}_i$.

Canonical cross loadings are correlations between the variables and variates within the different groups. In other words, 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 the X - set are given by $R_{xx}\hat{c}_i\hat{\rho}_i^*$ and the cross loadings for Y - set are given by $R_{yy}\hat{d}_i\hat{\rho}_i^*$.

3.2 Canonical variate scores

The canonical variate scores of X - set and Y - set of variables from the ith canonical variate pair 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^*. The predicted value of Y_i^* is given in (2).

$$\hat{Y}_i^* = \rho_i(X_i^* - \hat{c}_i'\,\overline{X}_i) + \hat{d}_i'\,\overline{Y} \tag{2}$$

3.3 Tests of significance

In order to perform CCA, we first determine 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. [2]. The test statistic is Wilk's lambda Λ[3] where k=min(p,q) and l_i is the ith 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 $H_o : \rho_1^* = \rho_2^* = \ldots = \rho_p^*$ and we can conclude that the X – set and the Y – set are dependent.

$$\Lambda(p, n-1-q, q) = \prod_{i=1}^{k}(1-l_i) = \frac{|S_{yy}^{-1}S_{yx}S_{xx}^{-1}S_{xy}|}{|S_{yy}|} \tag{3}$$

4. Results

4.1 Exploratory Data Analysis

The summary statistics of the data set are shown in Table 1. The predictor set (X - set) of variables represents the school characteristics and includes the following variables: Teacher to student ratio (X_1), School size (X_2), Proportion of students taking biology (X_3), Proportion of students taking physics (X_4) and Proportion of students with mean scores above C+ (X_5). The criterion set (Y - set) of variables represents the level of performance in career subjects in STEM education and includes the following variables: Mathematics mean score (Y_1), Biology mean score (Y_2), Physics mean score (Y_3) and Chemistry mean score (Y_4). 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.

TABLE I. SUMMARY STATISTICS FOR KCSE EXAMINATION RESULTS IN STEM SUBJECTS AND SCHOOL CHARACTERISTICS DATA

Variables	N	Mean	Median	Minimum	Maximum	Std. Deviation
Predictor set						
X_1	77	0.2380	0.22	0.10	0.73	0.109
X_2	77	127.7100	115.00	22.00	339.00	74.923
X_3	77	0.8614	0.87	0.52	1.00	0.146
X_4	77	0.3320	0.27	0.04	1.00	0.213
X_5	77	0.3605	0.21	0.00	1.00	0.357
Criterion Set						
Y_1	77	4.3019	3.24	1.27	10.73	2.783
Y_2	77	5.0953	4.40	1.82	11.21	2.346
Y_3	77	5.1027	4.57	1.40	10.30	2.332
Y_4	77	4.8628	4.14	1.55	11.06	2.437

4.2 Correlations

The correlations between all the variables in the study are shown in Table 2. Most correlations are significant at the 0.01 level. The teacher to student ratio correlates negatively (-0.394) with the school size. 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 an increase in school size. The school size correlates highly with all variables except the proportion of students taking biology. 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 is highly correlated with the mean score in chemistry. This shows that higher proportions of students taking physics correspond to better scores in chemistry. The proportion of students with mean scores above C+ is very highly correlated with the mathematics mean score. The mathematics mean score and biology mean score correlate very highly with the mean score in chemistry. The physics mean score correlate very highly with the proportion of students taking biology and the chemistry mean score.

TABLE II. CORRELATIONS WITHIN AND BETWEEN THE PREDICTOR AND CRITERIONSETS 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).

4.3 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 3. Pillai's, Helling's, Wilk's and Roy's tests all confirm that at least one variate pair is significant with $p < 0.05$.

TABLE III. MULTIVARIATE TESTS OF SIGNIFICANCE

Test Name	Value	Approx. F	Hypoth. DF	Error DF	Sig. of F
Pillais	1.46726	8.22629	20	284	0.000
Hotellings	23.6396	78.6015	20	266	0.000
Wilks	0.0227	24.1304	20	226.48	0.000
Roys	0.95814				

4.4 Eigenvalues and Canonical Correlations

The canonical correlation coefficients and the eigenvalues of the canonical roots are reported in Table 4. The first eigenvalue is 0.95814 and its corresponding canonical correlation coefficient estimate is 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 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 $p < 0.05$ and thus their corresponding canonical correlation coefficients and eigenvalues are not interpreted.

TABLE IV. 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

4.5 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 5.

TABLE V. RAW AND STANDARDIZED CANONICAL WEIGHTS

	Variate 1		Variate 2	
	Raw canonical weights	Standardized canonical weights	Raw canonical weights	Standardized canonical weights
Predictor set				
X_1	-0.03319	-0.00361	-3.12329	-0.34021
X_2	0.00102	0.07638	0.00083	0.06201
X_3	1.03985	0.15201	4.12889	0.6036
X_4	0.99701	0.21264	-3.01848	-0.64379
X_5	2.09484	0.74749	0.38857	0.13865
Criterion set				
Y_1	0.17336	0.48242	-0.44137	-1.22826
Y_2	0.08885	0.20844	-0.01315	-0.03086
Y_3	0.06253	0.14582	1.28775	3.00309
Y_4	0.07734	0.18845	-0.67646	-1.64821

4.6 Canonical loadings

The canonical loadings are correlations between variable scores and variables in the same domain. In Table 6 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. The rest of the school factors correlate positively. The variable with the largest loading is the proportion of students with C+ and above. 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. For the second variate, the X - set variables with the largest loadings are the proportion of students taking biology and the proportion of students taking physics. 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.

TABLE VI. CANONICAL LOADINGS AND CROSS LOADINGS

	Canonical Loadings		Canonical Cross Loadings	
Variables	Variate		Variate	
	1	2	1	2
Predictor set				
X_1	-0.17084	-0.35606	-0.16722	-0.21724
X_2	0.84178	0.05465	0.82397	0.03334
X_3	0.3337	0.7743	0.32664	0.47241
X_4	0.69637	-0.63077	0.68164	-0.38484
X_5	0.98501	0.01467	0.96418	0.00895
Criterion set				
Y_1	0.98894	-0.05801	0.96802	-0.03539
Y_2	0.95308	-0.03648	0.93292	-0.02225
Y_3	0.95363	0.28957	0.93346	0.17667
Y_4	0.98274	-0.03521	0.96196	-0.02148

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.

4.7 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 6. 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. 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 and those taking physics 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.

4.8 Prediction

The two sets of variate scores obtained in from Table 5 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 of X_1^* can be used to predict Y_1^*, where the predicted score is given by (4).

$$\hat{Y}_1^* = \rho_1 (X_1^* - \hat{c}_1' \overline{X}_1) + \hat{d}_1' \overline{Y} \tag{4}$$

Similarly, the score of X_2^* can be used to predict Y_2^* .From the data, the predicted values of Y_1^* and Y_2^* are given in (5) and (6).

$$\hat{Y}_1^* = -0.00353X_1 + 0.07476X_2 + 0.14879X_3 + 0.20813X_4 + 0.73164X_5 - 5.21 \tag{5}$$

$$\hat{Y}_2^* = -0.20756X_1 + 0.03783X_2 + 0.36826X_3 - 0.39278X_4 + 0.08459X_5 - 3.1318 \tag{6}$$

From (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 $S_{yy}^{-1}S_{yx}S_{xx}^{-1}S_{xy}$. From (6) the variables that contribute 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 $S_{yy}^{-1}S_{yx}S_{xx}^{-1}S_{xy}$.

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 and Recommendations

Based on the standardized 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. 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

[1] 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.

[2] Alvin, C Rencher (2002). "Methods of multivariate analysis". In: Wiley Interscience.

[3] Bhuyan, KC (2005). Multivariate Analysis & Its Applications. New Central Book Agency.

[4] Bickel, Robert (2007). Multilevel analysis for applied research: it's just regression! Guilford Press.

[5] Bryant, Mykeko (2012). Cracking the code. Vol. 42. 8, pp. 16–17. isbn: 9789231002335.

[6] 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.

[7] Gonzalez, Heather B and Jeffrey J Kuenzi (2012). "Science, technology, engineering, and mathematics (STEM) education: A primer". In: Congressional Research Service, Library of Congress.

[8] Huyer, Sophia (2015). "Is the gender gap narrowing in science and engineering". In: UNESCO science report: towards 2030, p. 85.

[9] 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.

[10] 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.

[11] Nagel, Sarah (2017). "Policy Analysis on UNESCO's Action Plan for Gender Equality 2014-2021".

[12] Patton, Wendy and Mary McMahon (2014). Career development and systems theory: Connecting theory and practice. Vol. 2. Springer.

[13] Solon, Gary, Steven J Haider, and JeffreyMWooldridge (2015). "What are we weighting for?" In: Journal of Human resources 50.2, pp. 301–316.

[14] The World Bank (2016). World Development Report 2016: Digital Dividends. Vol. 65. 3, pp. 461–468. isbn: 978-1-4648-0671-1.

[15] Thompson, Bruce (1991). "A primer on the logic and use of canonical correlation analysis." In: Measurement and Evaluation in Counseling and Development.

[16] 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. 39