CHAPTER FOUR

RESEARCH METHODOLOGY

4.1 Introduction

This chapter consists of five sections. Section 4.2 gives a brief discussion about the data sources while Section 4.3 contains description of the variables used in this study. Section 4.4 and Section 4.5 discuss the information about the computer programmes used and a summary of the statistical procedures used in this study respectively. Finally, Section 4.6 concludes the entire methodology used in this study.

In this chapter, we attempt to estimate separately the parameters of a joint educational production function for urban and rural schools as well as nationally based on canonical correlation analysis. This technique utilizes the canonical variables to yield parameter estimates in term of the original economic variables (Vinod, 1968). Our main focus is to reveal the relationship between inputs and outputs for Malaysian educational production function by calculating its marginal rate of transformation between outputs, marginal elasticity of output with respect to input, marginal product, and marginal rates of substitution between inputs.

4.2 Data Description

In this study, we choose the school as the unit of analysis for the estimation of educational production function. The analysis is conducted using cross sectional data of school average performance from the IEA Third International Mathematics and Science
Study – Repeated (TIMSS-R) 1999\(^1\) for eighth grade students.\(^2\) The basic sample design for TIMSS 1999 is generally referred as a two-stage stratified cluster sample design. The first stage consisted of a sample of schools, which were stratified. The school sample-selection method used for first-stage sampling was based on a systematic probability-proportional-to-size (PPS) technique. The probability of selection for a school was proportional to the number of eighth-grade students in the school. The second stage consisted of a single classroom selected at random from the target grade in sampled schools (Gonzalez and Miles, 2001).

Since the sample size for TIMSS was designed to allow for analyses at the school and classroom levels, at least 150 schools had to be selected from the target population. This was done so that to achieve better sampling precision. A sample of 150 schools results in 95 percent confidence limits for school-level and classroom level mean estimates that are precise to within plus or minus 16 percent of their standard deviations. Six instruments were developed: a mathematics test, a science test, a student background questionnaire, a mathematics teacher questionnaire, a science teacher questionnaire and a school questionnaire. The mathematics and science tests were designed to measure understanding of the eighth-grade students in these areas.

The school samples were selected using a simple random sampling method from all the secondary schools in Malaysia including publicly and privately Islamic secondary schools. School-level selection exclusions consisted private secondary schools, private Chinese secondary schools, international secondary schools, specials secondary schools

---

1 Data are available at http://timss.bc.edu
TIMSS is an educational research project conducted by the International Association for the Evaluation of Educational Achievement (IEA), to investigate student achievement in Mathematics and Science in about 40 countries around the world. It is designed to provide policy makers, educators, researchers, and practitioners with information about educational achievement and learning contexts.

2 Refer to Gonzalez and Miles (2001): IEA TIMSS 1999 User Guide for International Database, p.5-3, the eighth grader equivalent to Form 2 lower secondary student in Malaysia education system.
for the physically and mentally disabled and very small schools. The school sample
design is monitored by Statistics Canada as assigned. Then, a single classroom of Form
2 lower secondary students is chosen randomly from a number of Form 2 classes in each
selected school. Thus, there is only single classroom per school. Within each school, all
students in the sampled classrooms present on the day of test administration were
requested to complete the tests and accompanying student background questionnaire.

In Malaysia, the sample contains data for 150 schools with 5,713 students. However, there are only 5,577 out of the 5,713 selected students who have answered
both the student background questionnaire, mathematics and science tests.

There are a variety of ways to define rural depending on the geographic
information of the school’s location. There is no single recommended definition. Rather,
the choice of how to define rural depends on the nature of the analysis carried out. As
classified by the Department of Statistics, Malaysia (2000), urban areas are defined as
gazetted areas with their adjoining built-up areas which had a combined population of
10,000 or more at the time of the 1991 Population Census. Built-up areas are defined as
areas contiguous to a gazetted area and had at least 60 percent of their population (aged
10 years or more) engaged in non-agricultural activities and at least 30 percent of their
housing units having modern toilet facilities. Rural areas refer to the reminder of the
classified areas above.

For this analysis, the response of type of community of school variable from the
school questionnaire main survey under IEA TIMSS 1999 was explored in determining
the rural and urban schools. Rural schools are defined as those schools located outside
large urban centre as suggested by Cartwright and Allen (2002).
We find a few missing data for the teachers with more than five years experience variable. This item non-response is due to school principals infrequently leave this item blank on the questionnaire while others do not have teachers with more than five years experience. Besides, the infinite value of natural logs transformation of female variable for the boy schools forces us to leave boy schools out of analysis. After all, the usable data were obtained from 131 schools and 4,854 students after excluding the missing value and making necessary corrections. In our sample, 80 schools are located in rural area with 2,924 students and 51 urban schools with 1,930 students.

The question of what constitutes sufficient sample size for a multivariate analysis does not have a simple and direct answer. As suggested by Thorndike (1978), two rules of thumb are used to the sufficient sample size question. First, one informal guide is that there should be 10 subjects for each variable and probably adds 50 to this number to ensure sufficient sample size for small sets of variables. Thus the first rule of thumb is that $N \geq 10 (p + q) + 50$, where in our study, $p=2$ and $q=7$. A second and more stringent rule which is also a function of the number of variables is that $N$ should be equal to the square of the number of variables. This rule, $N = (p + q)^2 + 50$, means that our required sample size would rise rapidly as the sets become large.

In our study, the usable sample sizes, 131 schools, are sufficient for stable results at national level. However, it is quite unlikely that the sample sizes satisfy the rules for analyzing the differences at the rural-urban level. Nevertheless, we should proceed to analyze the data of as many subjects as possible, while holding out a group for cross-validation, and we should exercise extreme caution in our interpretations (Thorndike, 1978).
4.2.1 School Data

In our study, this sample was chosen because of the availability of very rich data on the school resources. This study uses the responses from the school questionnaire main survey under IEA TIMSS 1999. This school questionnaire is addressed to school principals and department heads who are asked to supply information about their schools. The data used for analysis are number of student in the school, number of staff and the school instructional time. Data on students in the school provides the total school enrolment, whereas the number of staff describes the school’s professional full and part-time staff and the percentage of teachers at the school for five or more years. As for the instructional time, it indicates the amount of instructional time scheduled for the target grade, according to the school’s academic calendar.

Besides, secondary data are also utilized to get a general and overall picture of Malaysia’s per capita grant on non-teaching recurrent costs. The Per Capita Grant is based on the number of eligible students enrolled at a school of the current calendar year. The main items included are student welfare costs (for example, counseling services), library facilities, maintenance and repairs, utilities, other type of materials and supplies and miscellaneous (Educational Planning and Research Division, Ministry of Education, Malaysia, 2003). This is shown in Table 4.1.
Table 4.1 Per capita grant on non-teaching recurrent costs

<table>
<thead>
<tr>
<th>Total school enrolment, X</th>
<th>Library facilities</th>
<th>Counseling services</th>
<th>Recurrent expenses and miscellaneous</th>
</tr>
</thead>
<tbody>
<tr>
<td>X ≤ 100</td>
<td>RM 1,500.00</td>
<td>RM 600.00</td>
<td>RM 5,000.00</td>
</tr>
<tr>
<td>101 ≤ X ≤ 500</td>
<td>First 100</td>
<td>RM 1,500.00 +</td>
<td>RM 600.00 +</td>
</tr>
<tr>
<td></td>
<td>X &gt; 100</td>
<td>RM 14.00/ student</td>
<td>RM 5.50/ student</td>
</tr>
<tr>
<td>501 ≤ X ≤ 1000</td>
<td>First 500</td>
<td>RM 7,086.00 +</td>
<td>RM 2,794.50 +</td>
</tr>
<tr>
<td></td>
<td>X &gt; 500</td>
<td>RM 13.00/ student</td>
<td>RM 5.00/ student</td>
</tr>
<tr>
<td>X ≥ 1001</td>
<td>First 1000</td>
<td>RM 13,573.00 +</td>
<td>RM 5,920.00 +</td>
</tr>
<tr>
<td></td>
<td>X &gt; 1000</td>
<td>RM 12.00/ student</td>
<td>RM 4.50/ student</td>
</tr>
</tbody>
</table>

Source: Educational Planning and Research Division, Ministry of Education, Malaysia, 2003

4.2.2 Student Data

Each student in the sampled class was asked to complete the TIMSS 1999 mathematics and science achievement tests and a student questionnaire. The student questionnaire sought information about the student demographics, student’s home background, attitudes and beliefs about mathematics and science, and experiences in mathematics and science classes.

To provide an educational context for interpreting the mathematics and science achievement results in this study, we use the detailed information from students about their home background and how they spend their time out of school. Specifically, one set of questions addresses home resources and support for academic achievement. Another examines how much out-of-school time students spend on their schoolwork. In an effort to summarize this information concisely and focus attention on educationally relevant support and practice, TIMSS has combined information from individual questions to
form an index that was more global and reliable than the component questions. According to their responses, students were placed in a "high", "medium", or "low" category.

Home Educational Resources Index (HERI) was derived from student's reports on the availability of the number of books in the home; educational aids in the home (computer, study, desk/table for own use, dictionary); and their parents' education. Students assigned to the high level of this index reported coming from homes with more than 100 books, with all three study aids, and where at least one parent finished university. Whereas, students assigned to the low level had 25 or fewer books in the home, not all three educational aids, and parents that had not completed secondary education. The remaining students were assigned to the medium level (Mullis et al, 2000).

One of the major ways that students can consolidate and extend classroom learning is to spend time out-of-school studying or doing homework in school subjects. To summarize the amount of time typically devoted to homework, TIMSS constructed an index of out-of-school study time (OSTI) that assigns students to a high, medium, or low level on the basis of the amount of time they reported studying or doing homework in mathematics, science, and other subjects. Number of hours based on: no time = 0; less than 1 hour = 0.5; 1-2 hours = 1.5; 3-5 hours = 4; more than 5 hours = 7. Students at the high level reported spending more than three hours each day out-of-school studying all subjects combined. Students at the medium level reported spending more than one hour but not more than three, while those at the low level reported less than one hour per day of out-of-school study (Mullis et al, 2000).
4.3 Measurement of the variables

As indicated in the preceding chapter, there are several categories of variables that may be responsible for the school performance of students. In this section, we are concerned with the definition and measurement of each variable. The variables used are distinguished into three groups of variables, which are school outputs, school inputs and environmental inputs in the educational production process.

4.3.1 School outputs

School outputs are measured by school average standardized test scores in mathematics (MATH) and science (SCIENDE) in the IEA TIMSS 1999 for Form 2 students. The school average standardized test scores are calculated by the mean score of standardized test score in mathematics and science from all participants in each sampled school. The Ministry of Education, Malaysia (2003b) defined the standardized test score, t-score, which is arrived at by multiplying the z-score\(^3\) by 10 and adding 50. The t-score is derived from the z-score and is therefore based on a normal distribution. The fact that a z-score is directly associated with a standard deviation unit makes it useful in further statistical calculation and interpretation. However, it is difficult to visualize a student’s performance given only a z-score. The t-score helps clarify the student’s position in the normal distribution by linearly scaling the z-score to fit the more familiar “0 to 100%; 50% average” paradigm. The mean is assigned a t-score of 50% and each unit of standard deviation is represented by a 10% increase or decrease from the mean.

\(^3\) The z-score is the number of standard deviation units a student’s raw score is above or below the mean. It expresses how far a score is from the mean in terms of standard deviation units. Thus, a z-score is obtained by taking the raw score for a student, subtracting the mean (average) of all student scores and dividing by the standard deviation of the student raw scores.

\[ Z = \frac{x - \bar{x}}{\sigma} \], where \( \sigma \) is the standard deviation for the student raw scores.
Consequently, the z-score can be thought of as the raw score percentage adjusted to fit the normal distribution.

4.3.2 School Inputs

The four variables we use to represent school inputs are per pupil non-teaching expenditures (PPNTE), pupil teacher ratio (PTR), teaching experience (TE) and instructional hours (INSHRS). The latter three variables are derived from data of the school background questionnaire supplied by the school principals under IEA TIMSS 1999.

The per pupil non-teaching expenditure (PPNTE) variable is derived from the information as shown in Table 4.1. It is obtained by averaging the total expenditures of library facilities, counseling services, other recurrent expenses and miscellaneous on the total school enrolment. According to Jacques and Brorsen (2002), test scores were positively related to expenditure on instruction and instructional support, and are negatively related to expenditure on student support, such as counseling and school administration. Thus, we expect a negative relationship between PPNTE variable and the student achievement.

Measure of the quantity of teacher resources is represented by PTR. This index is the ratio between the school size and the total number of teachers. The total number of teachers included full-time and part-time teachers. In this analysis, we calculate one full-time teacher as 100% and it is represented by one FTE (full-time equivalents); one part-time as 50% and it is represented by 0.5 FTE. For each school, the pupil teacher ratio is arrived at total school enrolment divided by the total number of teachers.
We explicitly introduce TE variable in our estimation as the percentage of teachers with more than five years experience to investigate its contribution to the schooling outputs. The variable is directly taken from the school data under TIMSS school questionnaire.

Furthermore, we consider the influence of instructional hours into the school performance by taking the INSHRS variable into our estimation. INSHRS variable is defined as the percentage of yearly school hours spent on instruction and it implies the effectiveness of schools in optimizing the school hours on instructional work. It is obtained by dividing the total yearly instructional hours excluding lunch breaks, study hall time, and after school activities by total school hours in a school year including lunch breaks, study hall time, and after school activities. Logically, more instructional hours in classroom does increase learning (Brown and Saks, 1987).

4.3.3 Environmental inputs

Education do not happened at school solely. To control for differences in the production environment, the environmental inputs such as home educational resources and out-of-school study time were included as exogenous factors production. These two variables have been found to consistently influence student performance. In this report, HER is defined as the percentage of students/participants with at least medium level in home educational resources index. Besides that, OST, the percentage of students/participants with at least medium level in out-of-school study time index is taking into the consideration to examine the relationship between time spent doing homework and school average performance in mathematics and science.
To further investigate the gender influence in Malaysian educational production process, we include the FEMALE variable - as the percentage of female students in class, into the analysis. We calculate this variable by dividing the total number of female students in Form 2 classes by the total number of Form 2 students (male and female) and multiply by 100 for each sampled school.

In summary, the definition of educational input variables and their expected relationship are shown on Table 4.2 and Table 4.3.

**Table 4.2 Definition of the variables used**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>School Outputs</strong></td>
<td></td>
</tr>
<tr>
<td>MATH ((Y_1))</td>
<td>School average standardized mathematics test scores</td>
</tr>
<tr>
<td>SCIENCE ((Y_2))</td>
<td>School average standardized science test scores</td>
</tr>
<tr>
<td><strong>School Inputs</strong></td>
<td></td>
</tr>
<tr>
<td>PPNTE ((X_1))</td>
<td>Per pupil non-teaching recurrent expenditure (RM)</td>
</tr>
<tr>
<td>PTR ((X_2))</td>
<td>Pupil teacher ratio</td>
</tr>
<tr>
<td>TE ((X_3))</td>
<td>Percentage of teachers with more than five years experience (%)</td>
</tr>
<tr>
<td>INSHRS ((X_4))</td>
<td>Percentage of yearly school hours spent on instruction (%)</td>
</tr>
<tr>
<td><strong>Environmental Inputs</strong></td>
<td></td>
</tr>
<tr>
<td>HER ((X_5))</td>
<td>Percentage of students with at least medium level in home educational resources index (%)</td>
</tr>
<tr>
<td>OST ((X_6))</td>
<td>Percentage of students with at least medium level in out-of-school study time index (%)</td>
</tr>
<tr>
<td>FEMALE ((X_7))</td>
<td>Percentage of female students in class (%)</td>
</tr>
</tbody>
</table>
### Table 4.3 Expected sign of the educational inputs

<table>
<thead>
<tr>
<th>Variables</th>
<th>Expected Sign</th>
<th>Rational</th>
</tr>
</thead>
<tbody>
<tr>
<td>School Inputs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PPNTES</td>
<td>-</td>
<td>Non-teaching expenditure does not contribute directly into learning.</td>
</tr>
<tr>
<td>PTR</td>
<td>-</td>
<td>Students learn less in a bigger class.</td>
</tr>
<tr>
<td>TE</td>
<td>+</td>
<td>Students learn more from experienced teachers.</td>
</tr>
<tr>
<td>INSRHRS</td>
<td>+</td>
<td>More instructional hours does increase learning.</td>
</tr>
<tr>
<td>Environmental Inputs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HER</td>
<td>+</td>
<td>Academic support in the home environment encourages learning.</td>
</tr>
<tr>
<td>OST</td>
<td>+</td>
<td>As the amount of time student spends on homework increase, better academic performance will achieve.</td>
</tr>
<tr>
<td>FEMALE</td>
<td>+</td>
<td>Females obtain better result than males.</td>
</tr>
</tbody>
</table>

#### 4.4 Computer Programmes Used

Most of the statistical analyses were done using the Statistical Package for Social Science (SPSS) version 10.0 and StatistiXL version 1.1, a new data analysis package that runs as an add-in to Windows versions of Microsoft Excel. The SPSS menu does not offer canonical correlation analysis. Therefore, we conduct canonical correlation analysis by using SPSS Syntax through MANOVA with DISCRIM subcommand. All the graphical presentations are also produced using SPSS and StatistiXL.

#### 4.5 Methodology Used

Data analysis is divided into three main subsections. They are exploratory data analysis, testing of mean difference and explanatory data analysis. For all the statistical tests carried out, the chosen level of significance (α) is 0.05.
4.5.1 Exploratory data analysis

Exploratory data analysis is used to disclose the pattern and the distribution of the data. It is focused more on visual presentation of the data. In this analysis, the diagrammatic presentation of data chosen to study the distribution for rural-urban difference in a variable is box-plot (Bryman and Cramer, 2001).

Box plot provides the five-number summary of a distribution. The five-number summary refers to first quartile ($Q_1$), median, third quartile ($Q_3$), minimum and maximum. The box comprises the middle 50 percent of observations, which is called the inter-quartile range ($Q_3 - Q_1$). Thus the lower end of box is the first quartile and the upper end is the third quartile. The line in the box is the median. The whiskers (the vertical line extending on both side of the box) indicate the range of minimum and maximum observations in the distribution excluding outliers, which are separately indicated.

In addition, box-plot provides information about the shape and dispersion of a distribution. For a left-skewed distribution (negatively skewed), there is a long upper whisker and the median rests in the lower end of the box. Most of the observations concentrate in the higher end of the scale. For right-skewed distribution (positively skewed), there is a long lower whisker and the median lies in the upper side of the box. Most of the observations cluster in the lower end of the scale. If the length of the upper whisker and the lower whisker is almost equal and the median divides the box into half, this is a symmetrical distribution.
Alternatively, skewness of a normal distribution can be established from the skewness statistic. If the distribution has no skewness, then the skewness statistic will be zero. If the distribution has a positive skew, then the skewness statistic will be positive. If the distribution has a negative skew then the skewness statistic will be negative. As for any statistic, the actual values of the skewness statistics rarely turn out to be exactly zero. Hence, we do this by setting up a 95% confidence interval around the skewness score. If the 95% confidence interval includes the value zero then we cannot reject the hypothesis that the distribution has no skewness. (Refer to http://www.uccs.edu/~lbecker/freq.htm).

### 4.5.2 Testing of mean difference

To verify the rural-urban differences observed in box-plot, the difference in mean is tested by using independent samples t-test. Before ascertaining the mean difference of the aspect studied among rural and urban groups of a variable, Levene test is carried out to test the homogeneity of variance. The violation of homogeneity of variance affects the result of F-test and t-test seriously.

\[ H_0: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_c^2 \]

\[ H_a: \text{Not all } \sigma_i^2 \text{ are equal} \]  \hspace{1cm} (4.1)

**Levene test statistics**

\[
F = \frac{\left( \sum_{i=1}^{c} n_i (\bar{X}_i - \bar{X}_y)^2 \right) / (c-1)}{\left( \sum_{i=1}^{c} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 \right) / (n-c)} \hspace{1cm} (4.2)
\]

where \( c \) is the number of groups in a variable. Therefore, \( c = 2 \) in our analysis. The degree of freedom for numerator is \( c-1 \) and degree of freedom for denominator is \( n - c \). If the null hypothesis is rejected, it may be concluded that the variances of groups in a variable differ significantly. In Levene test, we make the assumptions that all the group
populations are normally distributed and the samples from each population are random and independent.

To test the rural-urban differences in mean for a variable, a suitable independent samples t-test has to be chosen between the pooled-variance t-test and separate-variance t-test. For equal-variance groups, pooled-variance t-test is more suitable. For unequal-variance categories, separate-variance t-test is more appropriate (Jobson, 1991; Bryman and Cramer, 2001).

\[ H_0 : \mu_R = \mu_U \]
\[ H_a : \mu_R \neq \mu_U \]  \hspace{1cm} (4.3)

where the subscript R and U are refer to the rural and urban group respectively. If the null hypothesis is rejected, it may be concluded that there is evidence of a mean difference in the aspect studied for rural and urban samples. In this test, the tested variable is assumed to be random and independent from normal populations.

**Pooled-variance t-test statistics**

\[
t = \frac{\bar{X}_R - \bar{X}_U}{S_p \sqrt{\frac{1}{n_R} + \frac{1}{n_U}}} \]  \hspace{1cm} (4.4)

where the degree of freedom is \( n_R + n_U - 2 \) and \( S_p = \frac{(n_R - 1)S_R^2 + (n_U - 1)S_U^2}{n_R + n_U - 2} \).

**Separate-variance t-test statistics**

\[
t = \frac{\bar{X}_R - \bar{X}_U}{\sqrt{\frac{S_R^2}{n_R} + \frac{S_U^2}{n_U}}} \]  \hspace{1cm} (4.5)
where the degree of freedom is defined as \[
\frac{\left(\frac{S^2_R}{n_R} + \frac{S^2_{U_i}}{n_{U_i}}\right)^2}{\frac{S^2_R}{n_R} \left(\frac{S^2_{U_i}}{n_{U_i}}\right)^2 \frac{1}{n_R - 1} + \frac{1}{n_{U_i} - 1}}
\]

4.5.3 Simple Regression Analysis

The simple regression analysis is used to explain the one to one relationship between school outputs and educational inputs.

\[Y_j = \beta_0 + \sum_{i=1}^{n} \beta_k X_k + \mu\] (4.6)

whereby,

\[k = 1, 2, \ldots, n\]

\(Y_j\) = the \(j\)th observed value for the dependent variable (school outputs)

\(X_k\) = the known value for \(k\)th independent variable on the \(j\)th observation (school inputs or environmental inputs)

and \(\mu\) is a stochastic error term, intercept \(\beta_0\) and \(\beta_k\) are coefficients of explanatory variables to be estimated. The coefficient of determination, R square is a proxy for the goodness of fit of the model. A high R square indicates a good fit. The beta coefficient is used to measure the influence of the independent variables in a model.

4.5.4 Canonical Correlation Analysis

The main explanatory data analysis employed is Canonical Correlation Analysis. We employ Vinod’s (Vinod, 1968) adaptation of Hotelling’s canonical correlation analysis in order to perform the parameter estimate for joint educational production with Cobb-Douglas functional form. This approach has been applied to analyze educational
production by Chizmar and Zak (1984) and Gyimah-Brempong and Gyapong (1991). For the purpose of our study, we assume the existence of achievement gap in rural and urban schools. We make joint production estimates for each of the three samples: rural, urban and the combined group. Comparing different equations (Chizmar and Zak, 1984) may lend insight into rural-urban differences in the achievement of mathematics and science subjects.

4.5.4a Why choose Canonical Correlation Analysis?

Canonical correlation analysis (CCA) is a type of multivariate linear statistical analysis, first described by Hotelling (1935), which is used in a wide range of disciplines to analyze the relationships between multiple independent and multiple dependent variables. In other words, it is an exploratory technique to discover patterns of linear relationships between two sets of variables (Hair et al, 1987, 1998). CCA focuses on the correlation between a linear combination of the variables in one set and a linear combination of the data in another set. The pairs of linear combinations are called canonical variates (CV). Their correlations are called as canonical correlations, \( R_c \) and it measures the strength of association between the two sets of variables.

The technique attempts to concentrate a high dimensional relationship between the two sets of variables into few pairs of CV. The first pair of CV maximizes the correlation between a linear combination of one set and a linear combination of the other. A second pair of CV is uncorrelated with the first pair and maximizes the correlation between linear combinations of variables after the variance due to the first pair of CV has been removed. Discovery of pairs of CV continues either until no significant linkages between sets remain in the residual correlation matrices or until as many pairs of CV
have been defined as there are variables in the smaller set. Since each pair of CV is calculated from the residuals of the pair(s) extracted earlier, the resulting CVs are orthogonal.

The goal of CCA is to find the canonical weights (coefficients) $\alpha_1, \alpha_2, \alpha_3, \ldots \alpha_p$ to be applied to the $p$ Y variables and $\beta_1, \beta_2, \beta_3, \ldots \beta_q$ to be applied to the $q$ X variables in such a way that the correlation between CV$_{Y1}$ and CV$_{X1}$ is maximized. Although the numerical methods required for CCA are much more complex than those required for computing a bivariate correlation coefficient, CCA can be conceptually understood in terms familiar from bivariate analysis (Cliff, N., 1987). Figure 4.1 presents a summary on how CCA works in this study.

**Figure 4.1 Canonical Correlation Analysis Network**

![Diagram of Canonical Correlation Analysis](image)


"By adjusting the weights, $\alpha_p$ and $\beta_q$, we maximize correlation between Y and X."

There are three reasons why CCA is used in this study. As supported by Gyimah-Brempong and Gyapong (1991), CCA is used because it is able to handle more than one
dependent variable in a single equation as well as take into account possible non-causal inter-relationships among the dependent variables. Estimating separate equations for each output neglects relationships among the output, while estimating a simultaneous equation model assumes that the relationship among the dependent variable is causal. Secondly, the parameter estimate by CCA has proven to be consistent and unbiased in the presence of joint production. It does not require a priori functional form specification and inclusion of inappropriate variables or omission of relevant variables, does not leads to distortions (Vinod, 1968; Ruggiero, 1996 and 1998). Thirdly, CCA takes into consideration aspect of joint production technology, such as scope economies (Refer to www.natiomaster.com).4

4.5.4b Parameter Estimates

These applications begin, however, with the assumption that production is technically efficient. We specify the joint production function of a generalized Cobb-Douglas form, which can be written as:

\[ Y_1^{a_1} Y_2^{a_2} = \gamma X_1^{\beta_1} X_2^{\beta_2} X_3^{\beta_3} X_4^{\beta_4} X_5^{\beta_5} X_6^{\beta_6} X_7^{\beta_7} \mu \]  

whereby,

Y₁ = School average standardized mathematics test scores
Y₂ = School average standardized science test scores
X₁ = Per pupil non-teaching recurrent expenditure (RM)
X₂ = Pupil teacher ratio
X₃ = Percentage of teachers with more than five years experience (%)
X₄ = Percentage of yearly school hours spent on instruction (%)
X₅ = Percentage of students with at least medium level in home educational resources index (%)

4 Refer to www.natiomaster.com: Economics of scope are conceptually similar to economies of scale, which refer to efficiencies associated with increasing or deceasing the scope of marketing and distribution; changes in the number of different types of products. An example of economies of scope: as the library facilities are prepared and longer library hours used, more students can be benefited from the available facilities.
\(X_6 = \) Percentage of students with at least medium level in out-of-school study time index (%)

\(X_7 = \) Percentage of female students in class (%)

and \(\mu\) is a stochastic error term, \(\gamma\) and \(\alpha_p, \beta_q\) are coefficients to be estimated. Rewriting equation (4.7) using mnemonic notation and taking natural logs gives:

\[
\sum_{j=1}^{p} \alpha_j \ln Y_j = \sum_{k=1}^{q} \beta_k \ln X_k + \ln \gamma + \ln \mu
\]

(4.8)

where in our study, \(p=2\) and \(q=7\).

Following Ruggiero (1998), \(\gamma\) is an efficiency index \((0 < \gamma \leq 1)\) where \(\gamma = 1\) if and only if production is technically efficient. With this specification, it leads to non-convexity of the output set. Hence, \(\ln \gamma = \ln 1 = 0\) and equation (4.8) can be written as:

\[
\sum_{j=1}^{p} \alpha_j \ln Y_j = \sum_{k=1}^{q} \beta_k \ln X_k + \ln \mu
\]

(4.9)

Using CCA, we create two variates, \(U\) and \(V\), consisting of log-linear combinations of the outputs and inputs, respectively:

\[
U = \sum_{j=1}^{p} a_j \ln Y_j \quad \text{and} \quad V = \sum_{k=1}^{q} b_k \ln X_k
\]

(4.10)

The weight \(A = (a_1, a_2)\) and \(B = (b_1, \ldots, b_7)\) are chosen so that the correlation between variates \(U\) and \(V\) is maximized, i.e.

\[
\rho^* = \max_{A,B} \text{Corr}(U, V)
\]

(4.11)

leading to estimates \(A^* = (a_1^*, a_2^*)\) and \(B^* = (b_1^*, \ldots, b_7^*)\). Recognizing that

\[
U = \rho^* V
\]

(4.12)

If we denote \(A^*\) and \(B^*\) as the first canonical coefficients corresponding \(\rho^*\), the first canonical correlation, the Vinod (1969) has suggested that estimates in equation (4.9)
can be obtained by substituting equation (4.10) and (4.11); evaluating the resulting equation with $A^*$ and $B^*$ yields

$$\sum_{j=1}^{p} a_j^* \ln Y_j = \sum_{k=1}^{q} b_k^* \ln X_k$$

(4.13)

Following Chizmar and Zak (1984) and Gyimah-Brempong and Gyapong (1991), the marginal elasticity of output $j$ with respect to input $k$ in equation (4.13) is

$$\text{ME} \ (Y_j, X_k) \equiv \frac{\partial \ln Y_j}{\partial \ln X_k} \equiv \frac{\rho^* b_k^*}{a_j^*}$$

(4.14)

And, the corresponding marginal productivity evaluated at mean values is

$$\text{MP} \ (Y_j, X_k) \equiv \left( \frac{Y_j}{X_k} \right) \text{ME} (Y_j, X_k)$$

(4.15)

From the marginal products, we calculate marginal rates of technical substitution between inputs as:

---

3 Basically, the estimation of the marginal elasticity of output $Y$ with respect to a given input $X$ can be written in the form of a mathematical formula as

$$\text{ME} \ (Y, X) = \frac{\Delta Y^*}{\Delta X^*} \times 100\% = \frac{\Delta Y \times X}{\Delta X \times Y}$$

In the case of log-linear model, the slope coefficient measures the marginal elasticity of $Y$ with respect to $X$.

$$\text{ME} (Y, X) = \frac{\Delta \ln Y}{\Delta \ln X} = \frac{\frac{1}{Y} \Delta Y}{\frac{1}{X} \Delta X} = \frac{\Delta Y \times X}{\Delta X \times Y}$$

4 In general, the marginal product of input $X$ in the production of output $Y$ can be written as

$$\text{MP} \ (Y, X) = \text{change in } Y \div \text{change in } X$$

$$= \frac{\Delta Y}{\Delta X} \times \frac{X}{Y} \text{ME} (Y, X)$$

5 Mathematically, MRTS at a given point of the isoquant would be equal to the negative ratio of the MP of the two inputs at that point. Refer to Pindyck and Rubinfeld (1989). *Microeconomics*, 2nd edn. Singapore: Macmillan. (p.182) for detail of expanding on this equation.

MRTS of input $X_1$ for input $X_2$, $X_1$ = - Change in $X_2$ / Change in $X_1$

$$= - \frac{\Delta X_2}{\Delta X_1}$$

or alternatively $= \frac{MP_{X_1}}{MP_{X_2}}$
\[ MRTS (X_1, X_2) = \frac{MP_{X_2}}{MP_{X_1}} \]  
(4.16)

Finally, the marginal rate of output transformation between the mathematics (\(Y_1\)) and science (\(Y_2\))^8 is

\[ MRT (Y_1, Y_2) = \frac{\partial Y_2}{\partial Y_1} = \frac{a_1^* Y_2}{a_2^* Y_1} \]  
(4.17)

### 4.5.4c Sensitivity Analysis

Sensitivity analysis is conducted to ensure that the results are valid and can be generalized to the population. This approach is to assess the sensitivity of the results to the removal of a independent variable. Because the canonical correlation procedure maximizes the correlation and does not optimize the interpretability, the canonical weight and loadings may vary substantially if one variable is removed. To ensure the stability of the canonical weights and loading, the researcher should estimate multiple canonical correlations, each time removing a different independent variable (Hair et al, 1998).

### 4.5.4d Practical Issues

The following discussion provides only a list of the most important assumptions of canonical correlation analysis, and the major threats to the reliability and validity of results (Tabachnick and Fidell, 1983).

---

^8 MRT of output \(Y_1\) for output \(Y_2\) = Change in \(Y_2\)/ Change in \(Y_1\) = \(\frac{\Delta Y_2}{\Delta Y_1}\)
Normality

The tests of significance of the canonical correlations are based on the assumption that the distributions of the variables in the population (from which the sample was drawn) are multivariate normal. Multivariate normality is a generalization of the normal distribution to the joint distribution of two or more variables. Multivariate normality is not itself a testable hypothesis, but the likelihood of it is increased if the dependent variables (DVs) and independent variables (IVs) are all normally distributed.

In reality, there exists no graphic or convenient statistical procedure to test for multivariate normally. Therefore, most researchers test for univariate normality of each variable. Although univariate normality does not guarantee multivariate normality, the probability of multivariate normality for most real variables in social science is increased if all the variables have normal distributions (Hair et. al, 1998, Tabachnick and Fidell, 1983).

As suggested by Tabachnick and Fidell (1983), the normal and residuals plots is used to examine on the aspect of normality. In addition, the Kolmogorov-Smirnov one-sample test is used to support the normality result. This test is designed to measure whether a particular distribution differs significantly from a normal distribution (skewness and kurtosis of the distribution = 0).

We do not worry much on violating of this assumption as according to Ruggiero (1996, 1998), the canonical regression technique will provide consistent estimates of the slope parameters even without assuming multivariate normality.
Outliers

Outliers are observations with a unique combination of characteristics identifiable as distinctly different form the other observations. Outliers can be found in both univariate and multivariate situations and they can greatly affect the magnitudes of correlation coefficients. Graphically, multivariate outliers can be observed through a box-plot of standardized residuals, which there is a poor fit between obtained and predicted DV scores. Multivariate outliers will produce points outside the general swarm of points. (Refer to http://www.pfc.forestry.ca/profiles/wulder/mvstats/outliers_e.html).

Multicollinearity

Multicollinearity is a condition in which a set of predictor variables (DV's) are highly correlated among themselves and will confound the ability of the technique to isolate the impact of any single variable, making interpretation less reliable. In order to assess the degree of multicollinearity among IV's, we generate and then examine the Pearson product-moment correlation coefficient matrices between variables. To shed further light on the nature of multicollinearity problem, a series of Auxiliary Regression with 95% confidence level are estimated to examine the relationship between the independent variables and Variance inflation factor (VIF) is used to assess multiple collinearity. High VIF (>10.00) indicates serious multicollinearity (Gujarati, 2003).

Linearity

According to Tabachnick and Fidell (1983), linearity is important to canonical analysis in at least two ways. The first is that the analysis is performed on a set
correlation matrices that contain Pearson product-moment correlations. Pearson $r$ is sensitive to the straight-line component of relationships but not to other component. Residuals plots can be first be examined for violation of linearity between any variables. Linearity is also important to canonical correlation in that the analysis maximizes Pearson $r$ between a combination of variables from one set and a combination from the other. Therefore, inspection of scatterplots between all pairs of canonical variates may reveal departures from linearity. Referring to our model in Equation (4.9), the equation is linear in term of the logarithms of the variables. Thus, we expect linearity between pairs of canonical variates.

4.5.4e Test of Significance

According to Vinod (1968), Canonical correlation's significance is tested via a Bartlett's Chi-square and C.R.Rao's F. To test the null hypothesis that the $p$ dependent variables, that is the educational outputs are uncorrelated to the $q$ educational inputs as the independent variables, that is,

$H_0: \rho^* = 0$

$H_1: \rho^* > 0$  \hspace{1cm} (4.18)

Bartlett's $\chi^2$ approximation for the distribution for Wilks' Lambda variable, $\Lambda$ is

$$\chi^2 = - [(N-1)-\frac{1}{2}(p+q+1)] \ln \Lambda \hspace{1cm} (4.19)$$

The null hypothesis is rejected if $\chi^2 > \chi^2_a$ with $pq$ degrees of freedom.

And, C.R. Rao's F:

$$F = \left(1 - \frac{\Lambda^{1/s}}{\Lambda^{1/s}}\right) \left(\frac{ms + 2v}{pq}\right) \hspace{1cm} (4.20)$$
where

$$m = [(N-1) - \frac{1}{2} (p+q+1)],$$

$$s = \left[ \frac{p^2 q^2 - 4}{p^2 + q^2 - 5} \right]^2$$

$$v = \left[ (-1) \left( \frac{pq - 2}{4} \right) \right]$$

and $N$ is the total number of observations. The degrees of freedom for $F$ will be, respectively, $(pq)$ for the numerator and $(ms + 2\nu)$ for the denominator. The standard tables for Chi-square and F can be used to test the significant of the canonical fit by using Vinod’s technique.

As suggested by Hair et al. (1998), a multivariate test of all canonical roots also can be used for evaluating the significance of canonical roots in order to separate tests of each canonical function. Many of the measures for assessing the significance including Wilks’ lambda, Pillai’s trace, Hotelling’s trace and Roy’s ger. Nevertheless, the practical significance of the canonical functions, represented by the size of the canonical correlations ($R_c$), also should be considered when deciding which functions to interpret. Furthermore, canonical correlation coefficient ($R_c$) is a measure of the strength of the relationship between two sets of canonical variates.

4.5.4f The Redundancy Analysis

The squared canonical correlation ($R_c^2$) refers to the shared variance explained in the canonical variates (linear composites). Although this is a simple and appealing measure of the shared variance, it may lead to some misinterpretation because the squared canonical correlations represent the variance shared by the canonical variates,
not the original variables. Stewart and Love (1968) observed that redundancy indexes, $R_d$ since they better summarize the overlap variance between two sets of variables than the squares of canonical correlations, $R^2_c$. It is quite possible for a pair of canonical variates that have a large squared correlation not to explain much of the variance in the variables, that is, the canonical analysis may produce a pair of highly correlated weighted combinations of the variables that extract only a very small amount of the variance in the original variables.

As such, the $R_d$ index provides a summary measure of the ability of a set of independent variables (taken as a set) to explain variation in the dependent variables (taken one at a time). This index serves as a measure of accounted-for variance, similar to the $R^2$ calculation used in multiple regression. The calculation of redundancy index of a variate is the shared variance of the variate multiplied by the square canonical correlation. The larger the redundancy indexes, the larger the overlap of the variables in each domain (Hair et al., 1998).

4.6 Summary

This study centered on the relationships between the educational outputs, particularly, mathematics and science test scores and school inputs after controlling environmental inputs for Malaysian schools. The use of the canonical correlation analysis was aimed at giving the parameter estimate for joint educational production with Cobb-Douglas functional form, as has been employed by Chizmar and Zak (1984) and Gyimah-Brempong and Gyapong (1991).