ANLY482 ANALYTICS PRACTICM

INTERIM REPORT

Making Better Subject Combination Choices for Students

Heng Kok Chin

Peh Zhan Hao

Tan Yong Kiong, Alson

Table of Contents

Introduction

Education in Singapore

As a small country, Singapore is heavily reliant on its population due to its resource constraints. Today, her students are key to shaping the future of Singapore because their skills and intellectual ability will determine the opportunities Singapore attract. Education in Singapore is managed by the Ministry of Education (MOE), which controls the development and administration of Government-funded schools, the Institute of Technical Education (ITE), polytechnics and universities.

Singapore’s education system is highly regarded by the international community. The country has one of the highest achieving students in international education examinations, with teenagers coming in top in mathematics, reading and science tests. As such, Singapore’s education system has been recognized to be ahead of those across Asia, Europe, North and South America (Coughlan, 2016).

Singapore offers many diverse education paths for all age groups and academic abilities, from primary school to university. After six years of education, all primary school students across Singapore will have to sit for the Primary School Leaving Examination (PSLE), where they will be tested in 4 main subjects: English, Mathematics, Mother Tongue and Science. Students will then choose secondary schools of their choice based on their results in the examination. Depending on their merit and order of choice, they will subsequently be assigned to a secondary school.

At the end of their second year in secondary school, students are required to choose their subject combinations for the Singapore-Cambridge General Certificate of Education (Ordinary Level) Examination (GCE ‘O’ Level), according to their academic performance and choices offered by their respective schools. The total number of subjects that they have to sit for at ‘O’ Level ranges between six and ten, with English, Mother Tongue or Higher Mother Tongue, Mathematics, one Science and one Humanities Elective being the compulsory subjects.

About Edgefield Secondary School

Edgefield Secondary School (ESS) is a neighbourhood secondary school located in the North-East region of Singapore, striving to provide quality education to students living around the estate.

Equipped with the newest facilities and the latest technologies coupled with curriculum innovation, the school is committed to providing the ideal learning environment and experiences for its students. Currently, the school is in the midst of setting up their own Data Analytics Team to analyse trends and tackle educational problems faced by teachers and students, so as to improve on their decision making.

Problem Recognition

In Singapore, the GCE ‘O’ Level examination is a major examination in the education journey of many students. Their academic performance in the examination will affect the future education paths and course options available to them, whether they can be admitted into a Polytechnic or a Junior College, and whether they qualify to take on a particular course. For example, if a student wants to study Pure Biology in a Junior College, it is a prerequisite for them to have studied either Pure Biology or Combined Biology for his or her GCE ‘O’ Level. In addition, there are different entry requirements for a student to take Pure Science in Junior College, depending on the courses he or she has taken previously. For example, to qualify for Pure Biology in Junior College, a student is only required to pass Pure Biology in GCE ‘O’ Levels. However, if he or she is taking Combined Science (Biology/Chemistry), he or she would need to score at least a B4 in order to qualify. Hence, it is clear that the choice of subject combination is a complex decision, and it is important for teachers and schools to aptly recommend students the right subject combination based on their capabilities.

However, the dilemma that many schools and teachers face is how to aptly adopt and establish the right criteria to offer the right subject combinations to students, so as to improve their learning outcomes. For instance, it is difficult for teachers to decide whether or not to recommend or allow students take on the subject combinations of Double Science or Triple Science. Should schools determine the capability of students based on their overall subject grades, or should they base their decisions on their individual subject grades (such as mathematics or science)? Often, many parents feel that their children are capable of coping with Double or Triple Science subject combinations, even if their secondary 2 examination grades shows otherwise. Without proper analytical evidence, it is difficult for teachers to convince parents that the recommended subject combination is the better option for their child.

Therefore, we strive to come up with an analysis that will help parents and students in choosing a subject combination that is both manageable (considering the student’s capability) and desirable (for the future of the student). Our analysis would enable teachers to help students make better choices, and prevent them from the situation where their future academic and career paths are constrained by his or her poor subject combination.

Project Motivation

As students ourselves who have undergone Singapore’s education system, we can clearly relate to this problem that students face. We understand the difficulties that teachers and students face in making this decision, and the consequences a student might face should he or she perform undesirably as a result of a wrong subject combination. Because of this, we are committed towards analyzing the historical trends and discover useful and actionable insights that would allow the school to make better decisions. In addition, as the school is currently in the preparations of setting up their data analytics team, our analysis and findings would be useful for the team to proceed with further analysis subsequently.

As such, we will propose an analytical model to shed light on a more scientific and data-driven approach for our Project Sponsor to formulate better streaming practices.

Project Objectives

Utilizing data of past students' grades from the school's database, we aim to discover useful and practical insights which will allow teachers to better decide and advise students on choosing their Secondary 2 subject combination, particularly on whether they should take one of the following subject combinations:

1. Combined Science;
2. 1 Pure Science and 1 Combined Science;
3. Double Pure Sciences; or
4. Triple Pure Sciences

We will attempt to analyse the trends of students' academic performance by examining their past subject grades and subject combinations.

To achieve the above mentioned, we will perform an in-depth analysis on the historical data with the following aims:

1. To help secondary schools and teachers better formulate the right streaming practices and criteria that would benefit all students; and
2. To develop an application using R for the school so that they can input future data to improve the accuracy of the model in predicting students' GCE 'O' Level examinations results.

Data and Tools Used

Data Source

Our project sponsor has provided us with 3 batches of historical data from the school's database, for students who took their GCE 'O' Level examinations and graduated in years 2014, 2015 and 2016. The data comprises of their respective subject results and grades for each of the continual and semester examinations (CA1, SA1, CA2 and SA2) from secondary 1 to 4, as well as their PSLE and GCE 'O' Level results. To protect the confidentiality of the students' identity, the school has coded the names of their students and provided us with their class and index numbers instead. The data will consist of the following:

Batch of 2014	Batch of 2015	Batch of 2016
Secondary 1 (2011)	Secondary 1 (2012)	Secondary 1 (2013)
Secondary 2 (2012)	Secondary 2 (2013)	Secondary 2 (2014)
Secondary 3 (2013)	Secondary 3 (2014)	Secondary 3 (2015)
Secondary 4 (2014)	Secondary 4 (2015)	Secondary 4 (2016)

And for each year, we are also given the breakdown of the various examinations that each student has to take in a year. Here is the breakdown of the various data for each year:

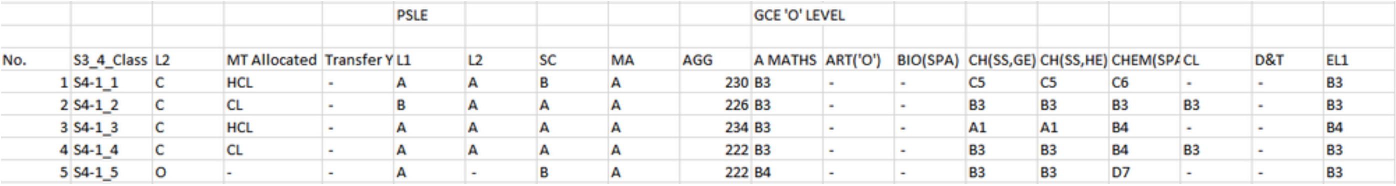
Secondary 1: CA1, SA1, CA2, SA2, Overall (5 sets of data)

Secondary 2: CA1, SA1, CA2, SA2, Overall (5 sets of data)

Secondary 3: CA1, SA1, CA2, SA2, Overall (5 sets of data)

Secondary 4: CA1 (Common Test), SA1, SA2 (Preliminary Examination), Overall (4 sets of data)

The 'Overall' refers to the overall score a student gets for that entire academic year. It is calculated by taking a combined score for CA1 & SA1 (37.5% CA1, 62.5% SA1) which makes up 40% of the total and CA2 & SA2 (25% CA2, 75% SA2) which makes up the remaining 60% of the total.

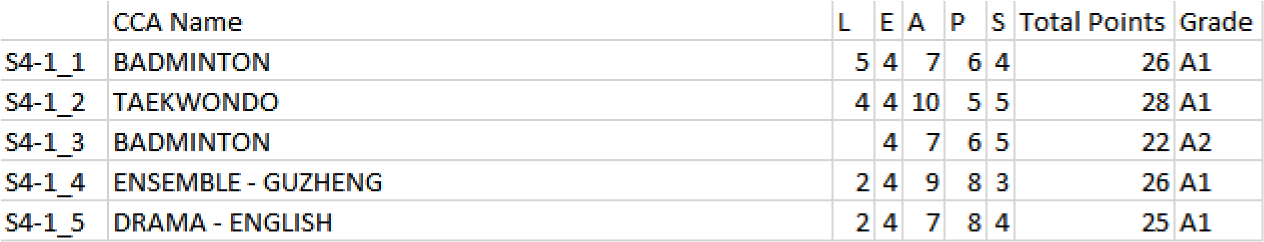


					PSLE					GCE 'O' LEVEL								
No.	S3_4_Class	L2	MT Allocated	Transfer Y	L1	L2	SC	MA	AGG	A MATHS	ART('O')	BIO(SPA)	CH(SS,GE)	CH(SS,HE)	CHEM(SPA	CL	D&T	EL1
1	S4-1_1	C	HCL	-	A	A	B	A	230	B3	-	-	C5	C5	C6	-	-	B3
2	S4-1_2	C	CL	-	B	A	A	A	226	B3	-	-	B3	B3	B3	B3	-	B3
3	S4-1_3	C	HCL	-	A	A	A	A	234	B3	-	-	A1	A1	B4	-	-	B4
4	S4-1_4	C	CL	-	A	A	A	A	222	B3	-	-	B3	B3	B4	B3	-	B3
5	S4-1_5	O	-	-	A	-	B	A	222	B4	-	-	B3	B3	D7	-	-	B3

This data is the first few columns of the Batch of 2016 CA1 data that we received. This file mainly contains the Secondary 1 CA1, Secondary 2 CA1, Secondary 3 CA1 and Secondary 4 CA1 results from the Batch of 2016.

Each individual student's name is being coded. For example, in the image shown, the first student is a Secondary 4 student from the class S4-1 and his index number is 1. This protects the identity of the students that we are analysing. Besides the main academic results, there are other columns such as the Second Language of the student, the results of PSLE and 'O' Levels (our main objective), the gender of the student and the student's class in Secondary 1 and Secondary 2 (for inter-class analysis). There are approximately 800 columns per file, and it varies based on the subjects offered during the particular year that the student is in.

Our team has obtained the CCA data of the students as well but only the CCA data during the students' graduating year. Here is a sample data of the CCA for the 'Batch of 2016':



	CCA Name	L	E	A	P	S	Total Points	Grade
S4-1_1	BADMINTON	5	4	7	6	4	26	A1
S4-1_2	TAEKWONDO	4	4	10	5	5	28	A1
S4-1_3	BADMINTON		4	7	6	5	22	A2
S4-1_4	ENSEMBLE - GUZHENG	2	4	9	8	3	26	A1
S4-1_5	DRAMA - ENGLISH	2	4	7	8	4	25	A1

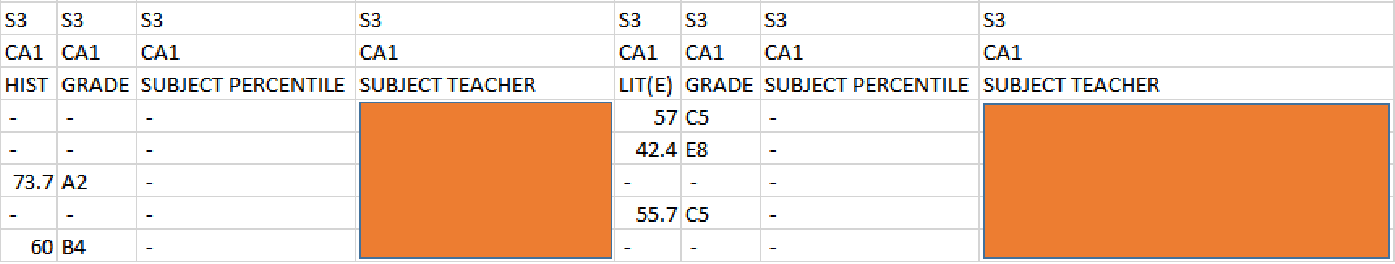
Our team was given the name of the CCA the student is involved in and also the number of points and the corresponding grade that the student received at the end of the four years of their secondary school. However, we are not given the CCA records at the end of each academic year.

Data Cleaning & Preparation

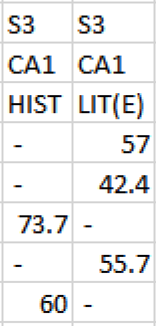
For our entire data cleaning, preparation and analysis, we will be using the following software:

1. Microsoft Excel
2. JMP Pro

In terms of initial data cleaning, we used Microsoft Excel as the dataset were given to us in that format. We have decided to use JMP Pro to perform our exploratory data analysis given its robustness and extensiveness. In addition, the capabilities of JMP Pro goes beyond exploratory analysis to provide more sophisticated predictive modelling, which we will be using to fulfill our objective of developing a predictive model for our project sponsor to adopt subsequently.



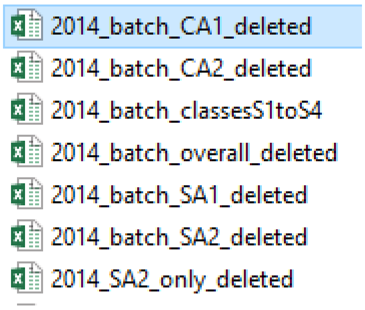
S3	S3	S3	S3	S3	S3	S3	S3
CA1	CA1	CA1	CA1	CA1	CA1	CA1	CA1
HIST	GRADE	SUBJECT PERCENTILE	SUBJECT TEACHER	LIT(E)	GRADE	SUBJECT PERCENTILE	SUBJECT TEACHER
-	-	-	[redacted]	57	C5	-	[redacted]
-	-	-	[redacted]	42.4	E8	-	[redacted]
73.7	A2	-	[redacted]	-	-	-	[redacted]
-	-	-	[redacted]	55.7	C5	-	[redacted]
60	B4	-	[redacted]	-	-	-	[redacted]



S3	S3
CA1	CA1
HIST	LIT(E)
-	57
-	42.4
73.7	-
-	55.7
60	-

Some columns are unnecessary and it will only add on to the size of the data and make things confusing. Such columns can be the grade of a particular subject for a student. The letter grade is derivable from the numerical score and thus we feel that it is unnecessary to keep the grade column. The name of the subject teacher is also unnecessary as we do not need to know the name of the teacher. Also, it is to protect the privacy of the teacher.

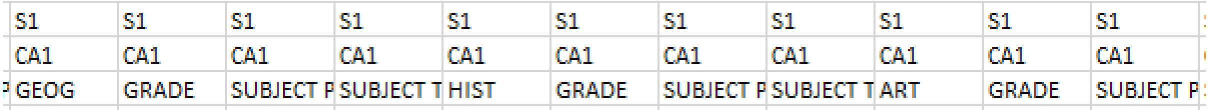
One other possible reason to remove a column can be that a particular subject is not being offered at all in that academic year. One of the signs of this occurance is that the data for that particular subject column are all empty. And after clarifying with our sponsor on which are the subjects not offered in the various academic years, we can safely remove those subject columns.



2014_batch_CA1_deleted
2014_batch_CA2_deleted
2014_batch_classesS1toS4
2014_batch_overall_deleted
2014_batch_SA1_deleted
2014_batch_SA2_deleted
2014_SA2_only_deleted

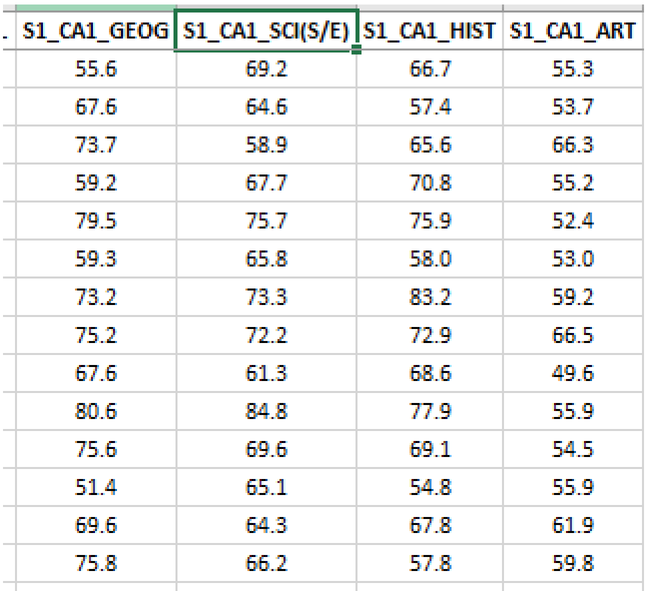
Figure 1: Data Transformation

Also, much effort and time was spent in the cleaning and preparation of our dataset. This was because the initial dataset given to us consisted of files for each examination for all Secondary levels i.e. 2014_Batch_CA1_deleted files consisted of CA1 examination scores for Secondary 1, 2, 3 and 4 (see Figure 2). Hence, if we had transformed the dataset into a chronological one i.e. Secondary 1 CA1, SA1, CA2, SA2, it would be complicated and time-consuming.



S1	S1	S1	S1	S1	S1	S1	S1	S1	S1	S1
CA1	CA1	CA1	CA1	CA1	CA1	CA1	CA1	CA1	CA1	CA1
GEOG	GRADE	SUBJECT P	SUBJECT T	HIST	GRADE	SUBJECT P	SUBJECT T	ART	GRADE	SUBJECT P

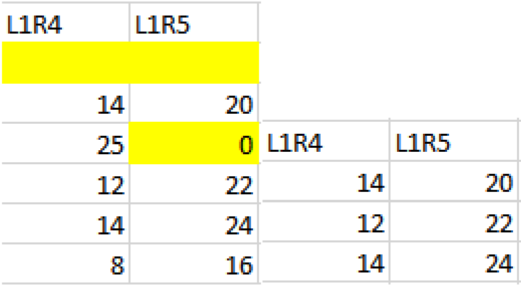
Figure 2: Before Exam Recoding



S1_CA1_GEOG	S1_CA1_SCI(S/E)	S1_CA1_HIST	S1_CA1_ART
55.6	69.2	66.7	55.3
67.6	64.6	57.4	53.7
73.7	58.9	65.6	66.3
59.2	67.7	70.8	55.2
79.5	75.7	75.9	52.4
59.3	65.8	58.0	53.0
73.2	73.3	83.2	59.2
75.2	72.2	72.9	66.5
67.6	61.3	68.6	49.6
80.6	84.8	77.9	55.9
75.6	69.6	69.1	54.5
51.4	65.1	54.8	55.9
69.6	64.3	67.8	61.9
75.8	66.2	57.8	59.8

Figure 3: After Exam Recoding

Given the nature of the dataset provided, we had to perform an extensive data cleaning and preparation step in order to ensure the consistency of our variables and to identify any inconsistencies and abnormalities. For example, we needed to recode each and every single subject and examination ID for the purposes of our analysis. In addition, the data provided consisted of a few students who retained and did not take their GCE ‘O’ Level examinations in the same year as his or her cohort, which resulted in missing data. As the subject combinations of the students differed from one another, we had to categorize them into their respective subject combination in order to proceed to our analysis.

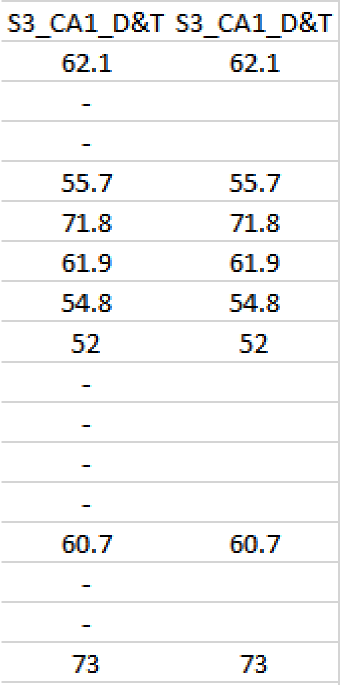


L1R4	L1R5
14	20
25	0
12	22
14	24
8	16

L1R4	L1R5
14	20
12	22
14	24

Figure 4: Before and After Removing of Students without ‘O’ Levels Results

As we require the GCE 'O' Levels L1R4 and L1R5 score for our analysis, any rows without this field will be removed. In addition, the data consisted of a few students who retained and did not take their GCE ‘O’ Levels in the same year as his or her cohort, which resulted in missing data. As such, to prevent skewing the results, we removed these unnecessary rows that we cannot make use of.



S3_CA1_D&T	S3_CA1_D&T
62.1	62.1
-	
-	
55.7	55.7
71.8	71.8
61.9	61.9
54.8	54.8
52	52
-	
-	
-	
-	
60.7	60.7
-	
-	
73	73

Figure 5: Removing Hyphens from Dataset

In JMP Pro 13, columns with hyphens will be treated as a nominal variable even though when a column is a numerical one (e.g. scores of subjects). As such, to make these columns appear as numerical variables so that we can use it to plot certain graphs, we need to replace the hyphens with blanks.

Data Methodology

Measuring and analysing the academic performance of students requires a multidimensional approach. Various past research have supported the hypothesis that the academic performance of students is dependent on each student’s socio-economic, psychological and environmental factors (Hijazi and Naqvi, 2006). For example, a student’s academic performance may be a result of tuition and family status and environment.

For the purposes of our analysis, we will focus our analysis solely on the environment factors which are directly attributed to our Project Sponsor, basing it on the context of Singapore as shown in Figure 7 below.

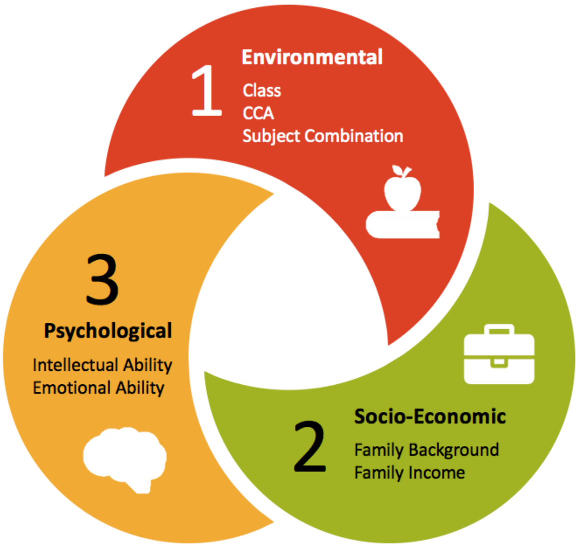


Figure 6: Data Methodology

According to a research conducted by the Bangladesh e-Journal of Sociology, there are three main variables in determining the academic performance of students, namely Environmental, Psychological and Socio-Economic factors. The environment that a student is immersed in school affects their academic performance, which includes their subject combination, class and co-curricular activities (CCA). As such, these factors form the basis of our analysis, which we have performed an exploratory data analysis on the historical results of the students.

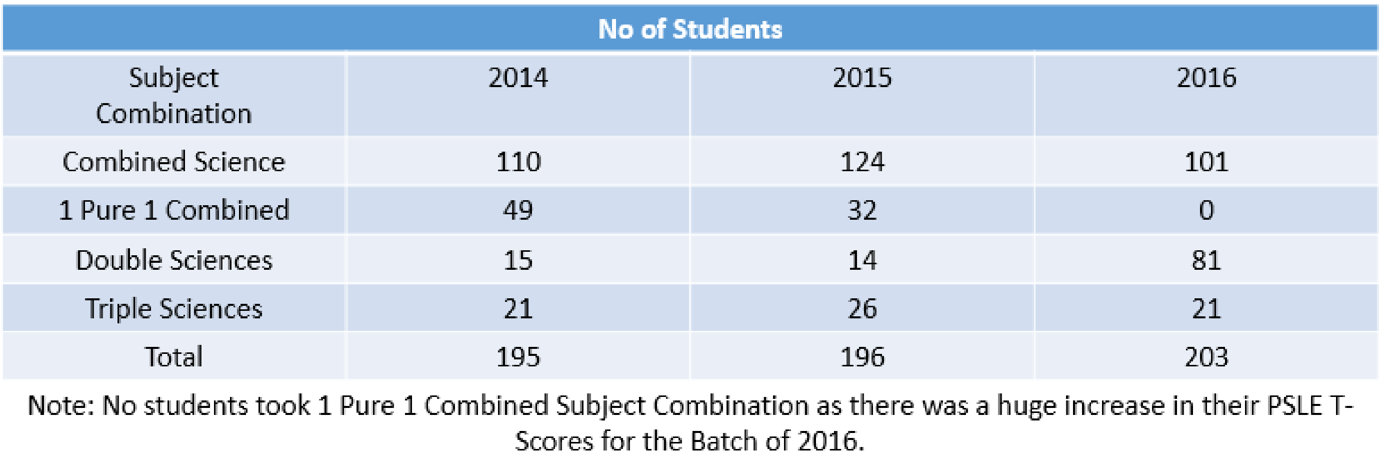
The psychological factors include their intellectual and emotional ability to cope with the workload. Finally, the socio-economic factors include their gender, race, family background, and parents' marital status.

Staffolani and Bratti (2002) asserted that the most important measure of the future academic performance and achievement of students is their previous educational outcomes, as cited from Ali, Shoukat, et al. (2013). In other words, the higher the student's past academic performance, the better their future academic performance. As such, we will be analyzing students' historical results which forms the basis of our analysis.

Exploratory Data Analysis

For our Exploratory Data Analysis (EDA), we performed some general descriptive statistics using JMP Pro to better understand the data before even venturing into analysing it. Below are some of the general descriptive statistics that we performed:

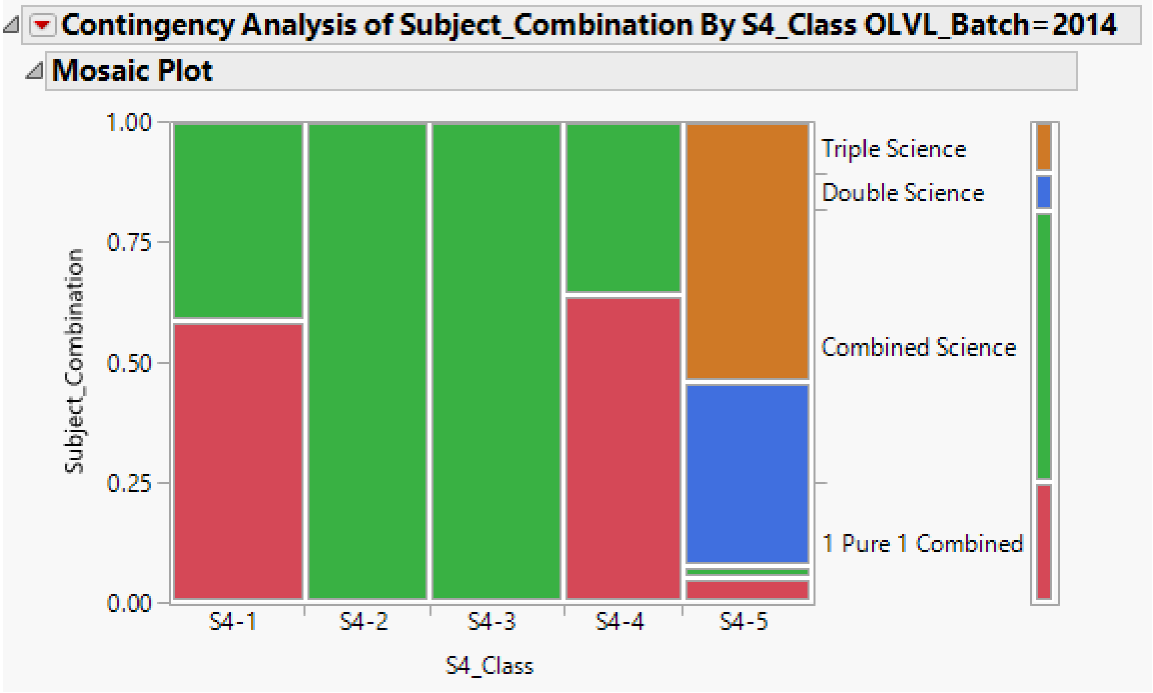
Composition of Students



No of Students			
Subject Combination	2014	2015	2016
Combined Science	110	124	101
1 Pure 1 Combined	49	32	0
Double Sciences	15	14	81
Triple Sciences	21	26	21
Total	195	196	203

Note: No students took 1 Pure 1 Combined Subject Combination as there was a huge increase in their PSLE T-Scores for the Batch of 2016.

Figure 7: Composition of Students by Subject Combinations



Contingency Analysis of Subject_Combination By S4_Class OLVL_Batch=2014

Mosaic Plot

Figure 8: Number of Students in each Subject Combination by Class - 2014 Batch

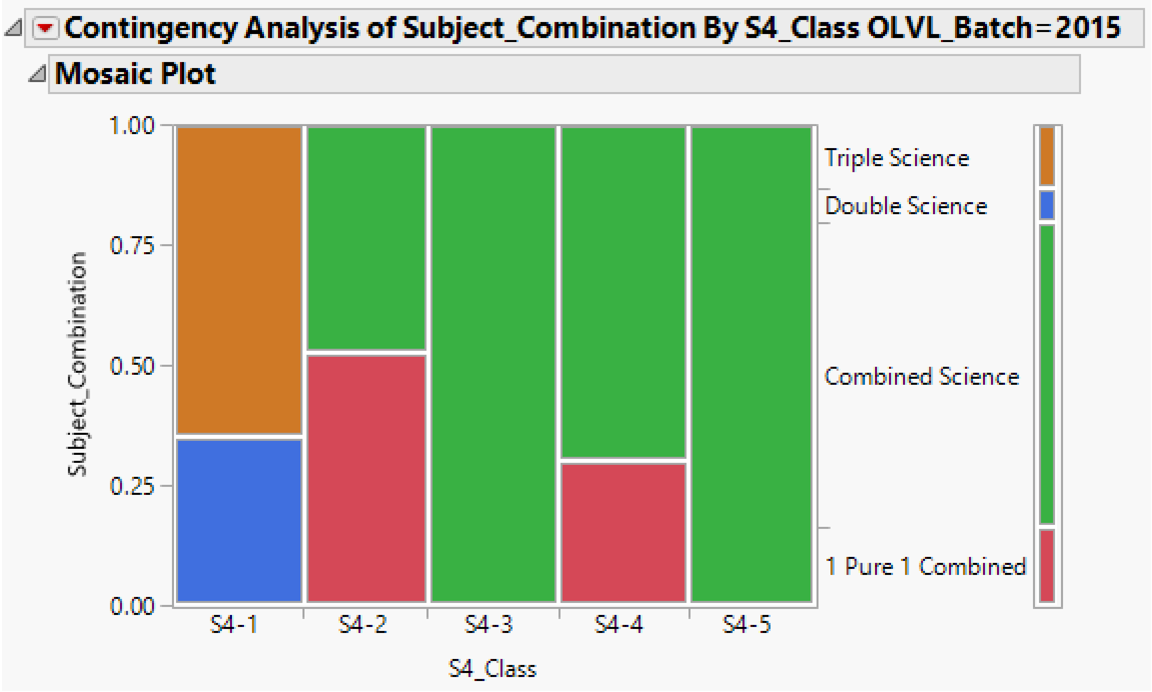


Figure 9: Number of Students in each Subject Combination by Class - 2015 Batch

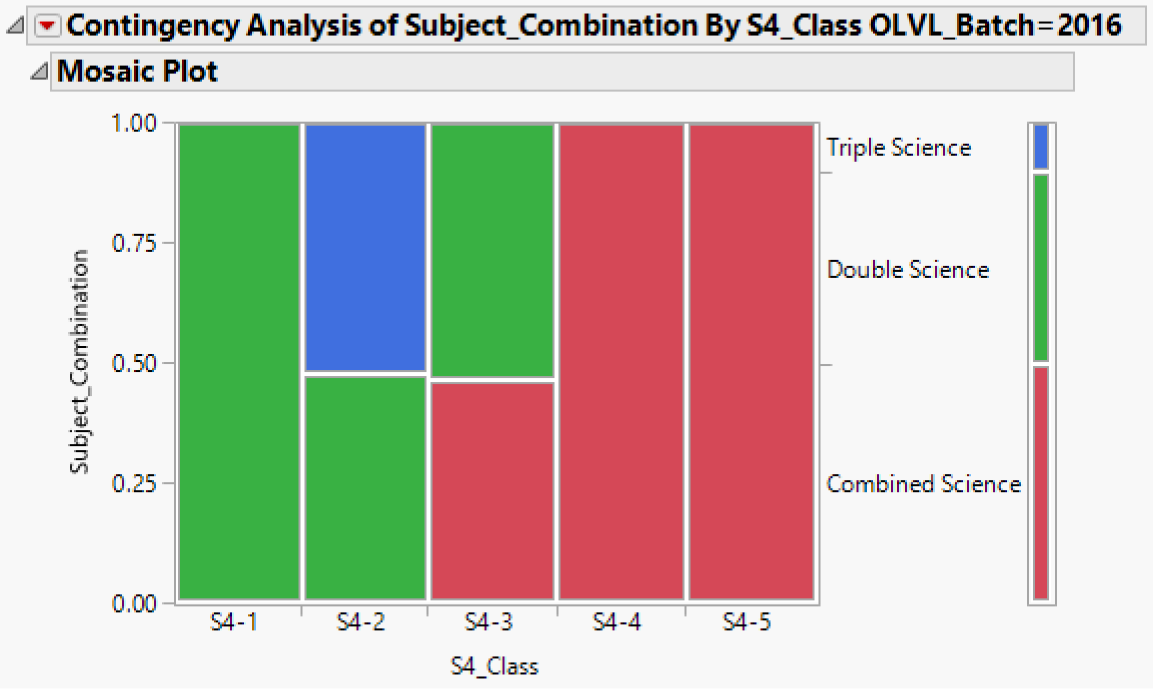
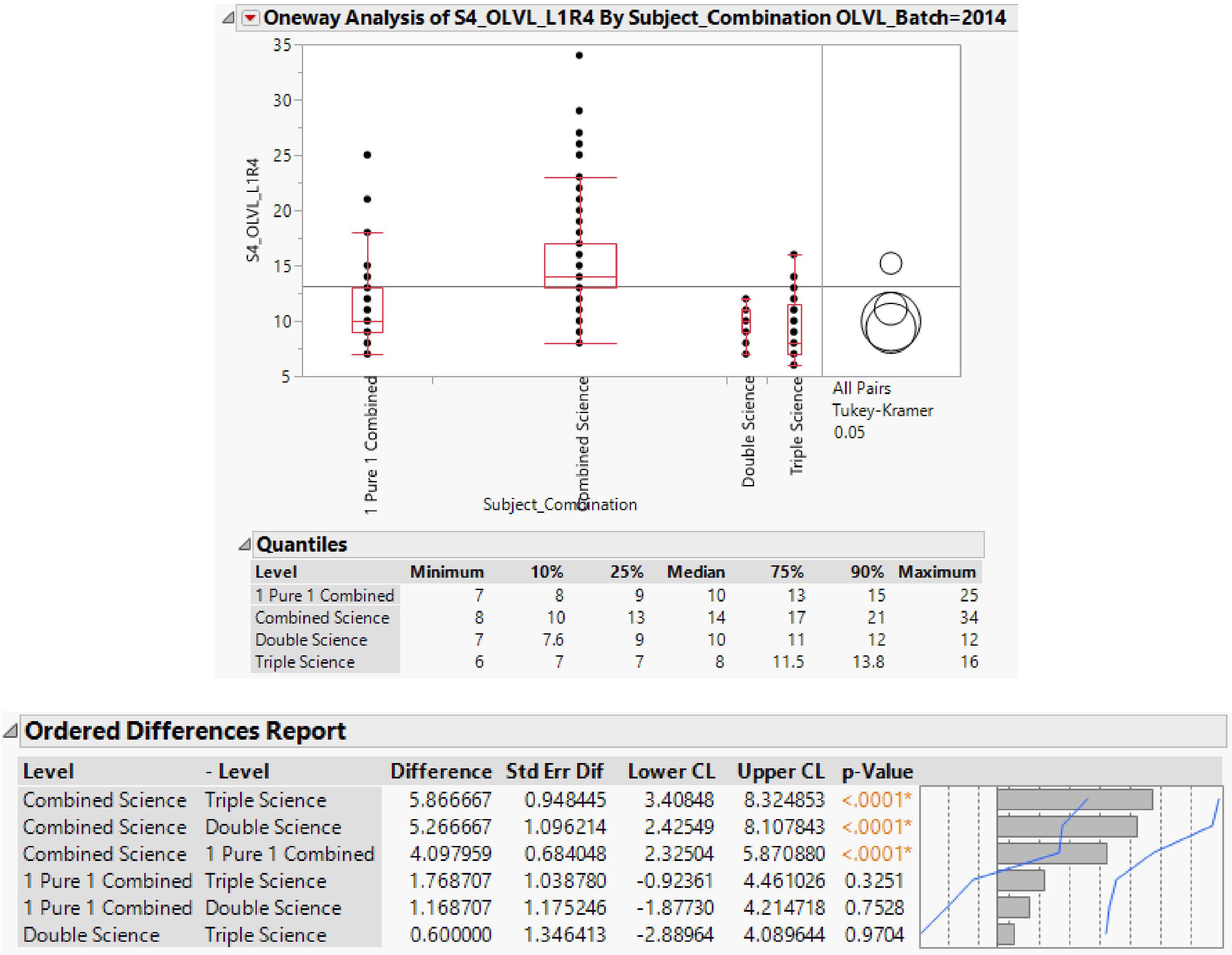


Figure 10: Number of Students in each Subject Combination by Class - 2016 Batch

'O' Levels Performance by Subject Combination

Subsequently, we checked on the 'O' Levels performance (L1R4 & L1R5) by subject combination for all 3 batches. Generally, the 3 batches exhibited similar trends, with students in 'Triple Science' subject combination performing better than those in 'Double Science', who in turn perform better than students in '1 Pure 1

Combined' and those in 'Combined Science'. We made use of Tukey's Range Test to determine significant differences of each of the pairs of subject combinations.



Oneway Analysis of S4_OLVL_L1R4 By Subject_Combination OLVL_Batch=2014

Quantiles

Level	Minimum	10%	25%	Median	75%	90%	Maximum
1 Pure 1 Combined	7	8	9	10	13	15	25
Combined Science	8	10	13	14	17	21	34
Double Science	7	7.6	9	10	11	12	12
Triple Science	6	7	7	8	11.5	13.8	16

Ordered Differences Report

Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
Combined Science	Triple Science	5.866667	0.948445	3.40848	8.324853	<.0001*
Combined Science	Double Science	5.266667	1.096214	2.42549	8.107843	<.0001*
Combined Science	1 Pure 1 Combined	4.097959	0.684048	2.32504	5.870880	<.0001*
1 Pure 1 Combined	Triple Science	1.768707	1.038780	-0.92361	4.461026	0.3251
1 Pure 1 Combined	Double Science	1.168707	1.175246	-1.87730	4.214718	0.7528
Double Science	Triple Science	0.600000	1.346413	-2.88964	4.089644	0.9704

Figure 11: Comparison of 'O' Level L1R4 - 2014 Batch

For the Batch of 2014, Combined Science students performed significantly worse than those in all other subject combinations in the 'O' Levels, having a p-value of less than 0.05 (see Figure 12).

Prelims & 'O' Levels Performance by Class

We also attempted to compare the Prelims and 'O' Levels performance by class to see if there is any deviation in trend. However, the general trend remains that the class with students taking the 'Triple Science' subject combination tends to do better than students in other classes taking other subject combinations (see Figure 13 and 14).

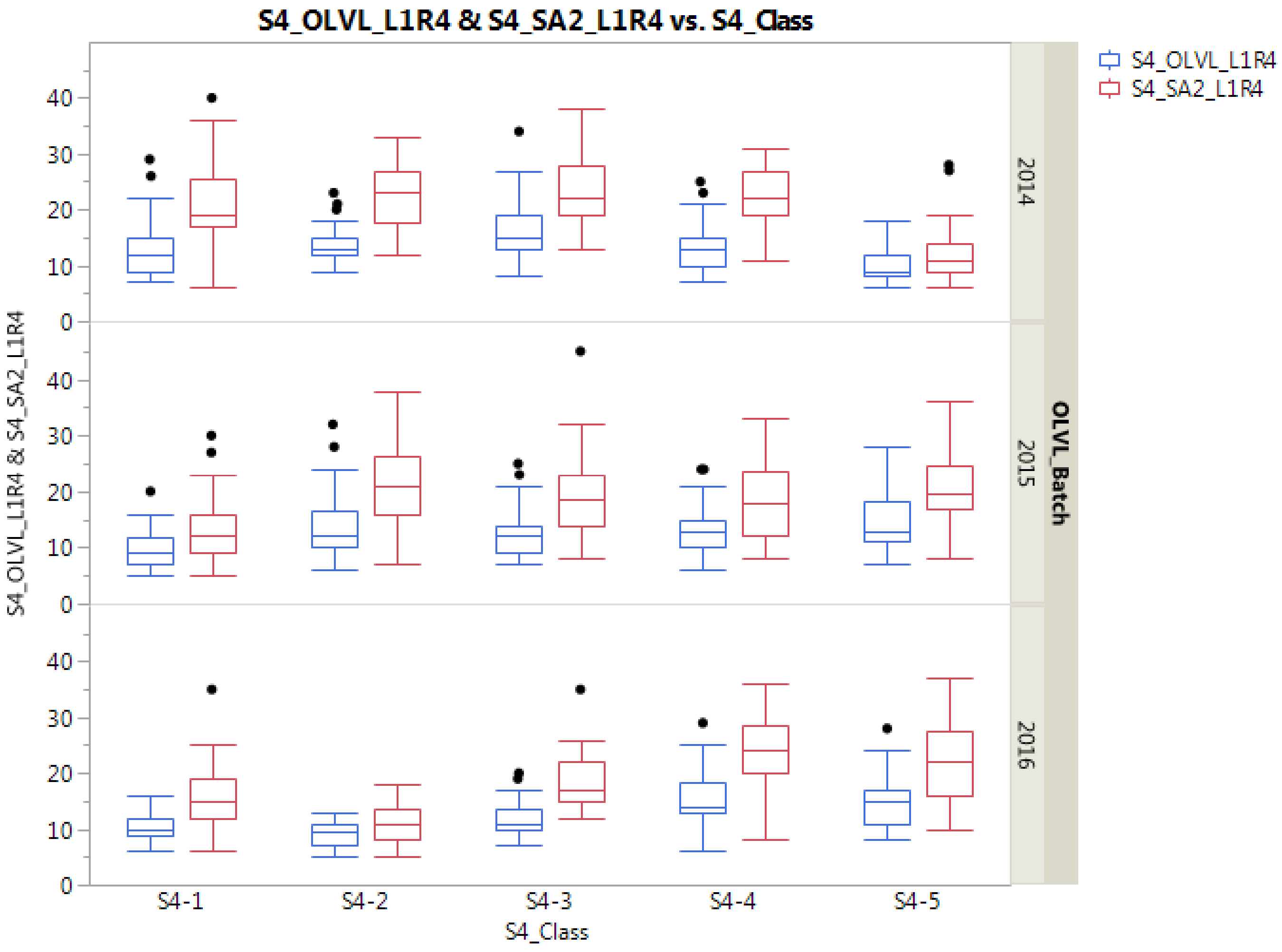


Figure 12: Prelims vs. 'O' Levels L1R4 side by side comparison

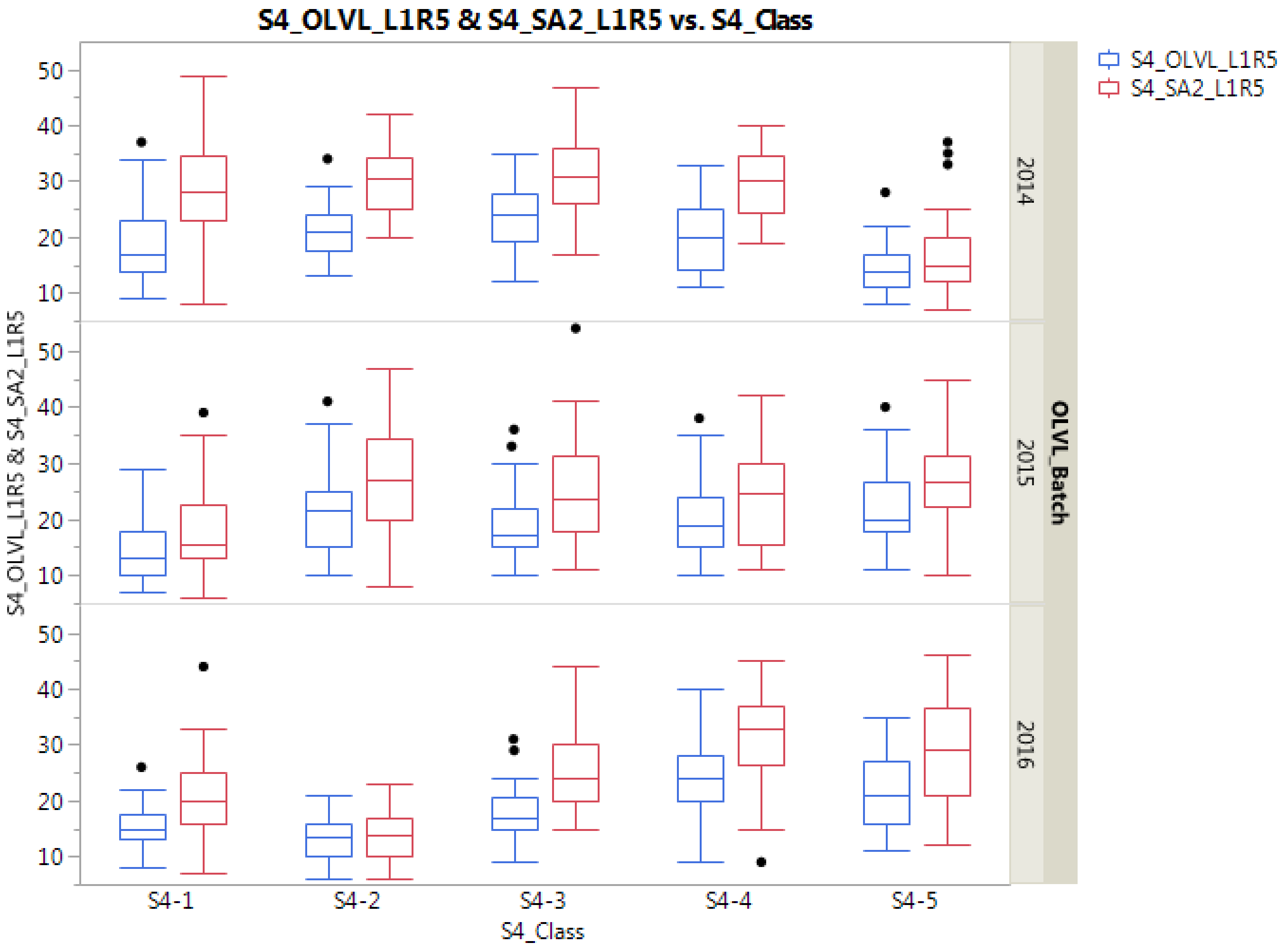
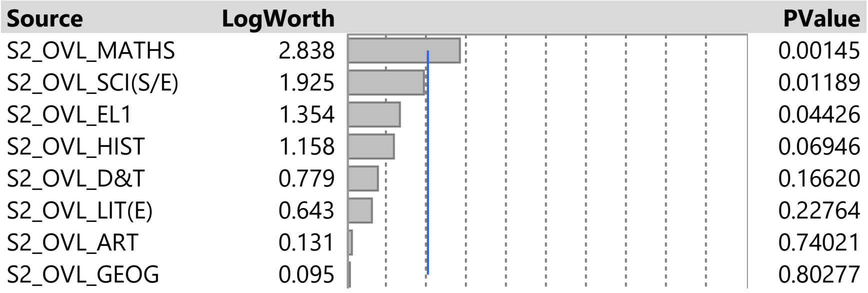


Figure 13: Prelims vs. 'O' Levels L1R5 side by side comparison

Evaluation of Current Practice

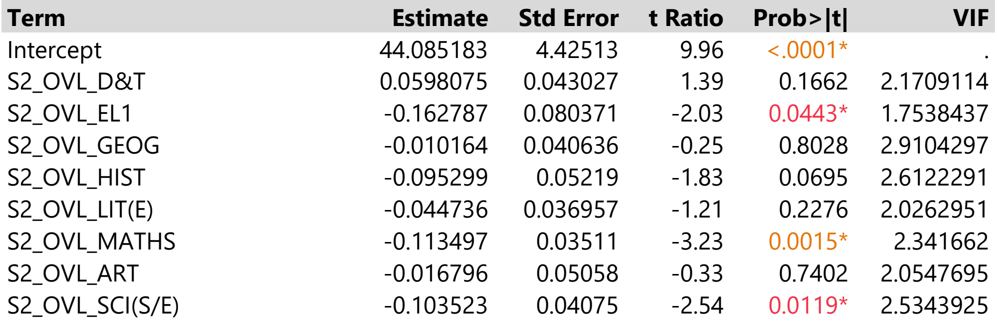
One of the objectives set out by our sponsor was that they wanted to find out whether Secondary 2 Mathematics scores or Secondary 2 Science scores is a better predictor of students' 'O' Levels performance, based on their subject combinations. The analysis below shows the Nominal Logistic Regression analysis for the data from the three respective batches with 'O' Levels L1R4/L1R5 versus the Secondary 2 individual subject scores. Firstly, both the Secondary 2 individual overall subject scores (independent variables) and L1R4/L1R5 scores respectively (dependent variable) are inserted into the Fit Model for analysis.



Source	LogWorth		PValue
S2_OVL_MATHS	2.838		0.00145
S2_OVL_SCI(S/E)	1.925		0.01189
S2_OVL_EL1	1.354		0.04426
S2_OVL_HIST	1.158		0.06946
S2_OVL_D&T	0.779		0.16620
S2_OVL_LIT(E)	0.643		0.22764
S2_OVL_ART	0.131		0.74021
S2_OVL_GEOG	0.095		0.80277

Table 1: Effect Summary for L1R4 Scores, Batch of 2014

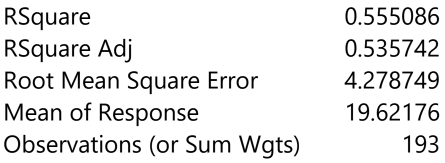
We consider Secondary 2 subjects with p-value less than 0.05 to be statistically significant in affecting the L1R4/L1R5 scores in the ‘O’ Levels. As seen from the ‘Prob>|t|’ column in Table 1 above for the ‘O’ Levels batch of 2014, the Secondary 2 results for Maths, Science and English are significant with p-value of less than 0.05. The Secondary 2 results for Maths, Science and English (in the order of importance) is statistically significant in impacting the L1R4 scores. For example, if a student’s Maths score increases by 1 units, the log odds of his L1R4 scores decreases by 2.838 units. The VIF is also evaluated to ensure that it does not exceed 8, which is a sign of multicollinearity. The VIF results in Table 1 above proves that each of the Secondary 2 subjects is not highly correlated with another subject.



Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	44.085183	4.42513	9.96	<.0001*	.
S2_OVL_D&T	0.0598075	0.043027	1.39	0.1662	2.1709114
S2_OVL_EL1	-0.162787	0.080371	-2.03	0.0443*	1.7538437
S2_OVL_GEOG	-0.010164	0.040636	-0.25	0.8028	2.9104297
S2_OVL_HIST	-0.095299	0.05219	-1.83	0.0695	2.6122291
S2_OVL_LIT(E)	-0.044736	0.036957	-1.21	0.2276	2.0262951
S2_OVL_MATHS	-0.113497	0.03511	-3.23	0.0015*	2.341662
S2_OVL_ART	-0.016796	0.05058	-0.33	0.7402	2.0547695
S2_OVL_SCI(S/E)	-0.103523	0.04075	-2.54	0.0119*	2.5343925

Table 2: Parameter Estimates for L1R4 Scores, Batch of 2014

As students obtain higher scores in their ‘O’ Level Examinations, they will obtain a lower L1R4/L1R5 results. Hence, the more negative the correlation of a particular Secondary 2 subjects versus the ‘O’ Level L1R4/L1R5 score, the more important is the Secondary 2 subject in determining the future ‘O’ Level results. Through Table 2, as the Secondary 2 results for Maths, Science and English increases, students tend to have lower L1R4 scores.



RSquare	0.555086
RSquare Adj	0.535742
Root Mean Square Error	4.278749
Mean of Response	19.62176
Observations (or Sum Wgts)	193

Table 3: Summary of Fit for L1R4 Scores, Batch of 2014

R-squared is a statistical measure of how close the data are to the fitted regression line. Through Table 3, we understand that 55% of the variability in the L1R4 score is explained by the mean.

Through similar analysis on the other two batches, we have discovered similar strong significance of the Secondary 2 Science, Maths and English results in predicting the L1R4 scores. However for the Batch of 2015, the order of significance are as follows: English, Maths and Science, while for the Batch of 2016, the order of significance on L1R4 scores are as follows: Science, English and Maths.

A summary of the statistical significance in illustrated in Table 4.

Statistical Significance on L1R4	2014	2015	2016
1	Maths	English	Science
2	Science	Maths	English
3	English	Science	Maths

Table 4: Statistical Significance of Secondary 2 Subjects on L1R4, all batches

However for the L1R5 scores, only the Science and Maths results are statistically significant for all batches, except the Batch of 2015 where English is considered a significant subject. The statistical significance of Secondary 2 subjects on L1R5 scores are shown in Table 5 below.

Statistical Significance on L1R4	2014	2015	2016
1	Maths	Maths	Science
2	Science	Science	Maths
3		English	

Table 5: Statistical Significance of Secondary 2 Subjects on L1R5, all batches

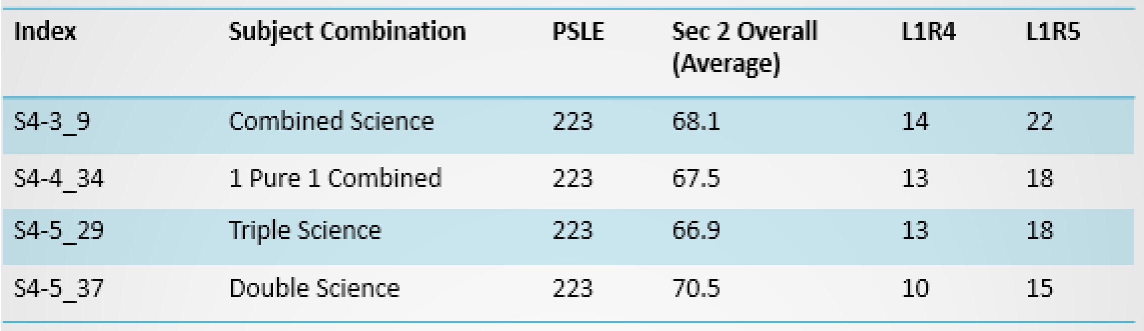
Hence, the school might also want to consider having English as an additional determining criterion for the students' subject combinations, as it has a significant impact on their 'O' Level results.

Time-series Analysis

To further analyze the performance of students, we selected a few students from each of the batch with similar PSLE scores and similar overall Secondary 2 scores. We drew overlay plots using JMP Pro 13 to see the Secondary 1 to Secondary 4 scores of these students. We want to see if these students, who ended up with different subject combinations, had any variations in their performance that is above or below the expectations of them.

This example is from the 'Batch of 2014': We chose the following students as shown in this table.

Table 6: Selected Students from 2014 Batch



Index	Subject Combination	PSLE	Sec 2 Overall (Average)	L1R4	L1R5
S4-3_9	Combined Science	223	68.1	14	22
S4-4_34	1 Pure 1 Combined	223	67.5	13	18
S4-5_29	Triple Science	223	66.9	13	18
S4-5_37	Double Science	223	70.5	10	15

Then here are the overlay plots of their results from Secondary 1 to Secondary 4, from top-left to top-right to bottom-left to bottom-right. The 'Exam' mentioned in the x-axis here refers to the CA1, SA1, CA2, SA2, Overall scores as described much earlier.

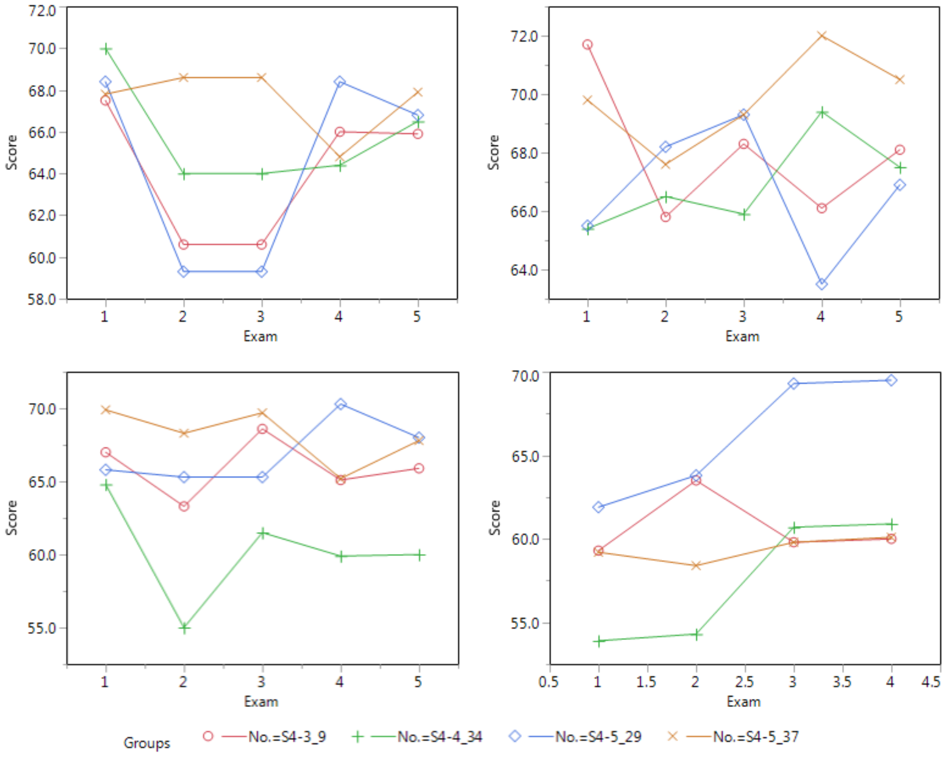


Figure 14: Time-series Analysis on Selected Students

What we can observe is that the 'Double Science' student here performed better than the 'Triple Science' student. The 'Triple Science' student performed equally well as the student taking '1 Pure 1 Combined' science subject combination. What we can draw is that students with similar PSLE scores and Secondary 2 overall scores who take different subject combinations can end up with very different scores.

Moving Forward

Following our analysis, we will continue to analyse deeper into the academic performance of individual students over time, and investigate the differences in academic performance (rate of change) of students in the various subject combinations. In other words, we seek to examine and identify if there are students who perform consistently better or worse (year-on-year) following their choice of subject combination after Secondary 2. Such analysis would offer insights as to whether or not the student is capable of coping with his or her subject combination.

We will also expand our analysis broader by analysing the other factors that could potentially affect the academic performance of students, such as their gender, race, parents’ occupation, parents’ marital status as well as socioeconomic status (SES). These factors can offer useful insights on the impact of a student’s family environment on their academic performance, which would allow the school to focus more attention to students in need.

Lastly, we will also develop a predictive model based on the historical data to project the academic performance of students for their GCE 'O' Level Examinations given their Secondary 2 individual subject results in their continual and semester examinations. This will greatly aid the school in recommending the right subject combinations to the students (and their parents) using analytical evidence and historical trends.

Stakeholders

The main stakeholders of this project include:

Project Supervisor

Prof Kam Tin Seong, Associate Professor of Information Systems; Senior Advisor, SIS (Programme in Analytics)

Sponsor

Mr Lee Peck Ping, Principal of Edgefield Secondary School (ESS)

Mrs Wong Puay Kheng, Vice Principal of ESS

Other Stakeholders

Students of ESS

Teachers and Heads of Department (HODs) of ESS

Parents of students studying in ESS

Project Members

Heng Kok Chin. Year 4 undergraduate from School of Information Systems

Peh Zhan Hao. Year 4 undergraduate from Lee Kong Chian School of Business.

Tan Yong Kiong. Year 4 undergraduate from Lee Kong Chian School of Business

References

Ali, S., Haider, Z., Munir, F., Khan, H., & Ahmed, A. (2013). Factors Contributing to the Students Academic Performance: A Case Study of Islamia University Sub-Campus. *American Journal of Educational Research, 1*(8), 283-289. doi:10.12691/education-1-8-3

Coughlan, S. (2016, December 6). *Pisa tests: Singapore top in global education rankings*. Retrieved from BBC: http://www.bbc.com/news/education-38212070

Hijazi, S. T., & Naqvi, R. S. (2006). Factors affecting student's performance: A Case of Private Colleges. *Bangladesh e-Journal of Sociology, 3*(1), 1-10.

Johnston, O., & Wildy, H. (2016). The effects of streaming in the secondary school on learning outcomes for Australian students – A review of the international literature. *Australian Journal of Education, 60*(1), 42-59. doi:10.1177/0004944115626522

Appendix

Metadata Table

Metadata	Type	Description
‘No.’	Nominal	Assigned unique value to differentiate between different students from different batches (goes from A001 to A196, B001 to B196, C001 to C207). A001 is first student in batch of 2014, B001 is first student in batch of 2015, C001 is first student in batch of 2016.
‘S4_Index’	Nominal	The class and index number of the students when they are in Secondary 4. Goes by the format S4-<classNumber>_<indexNumber>. Example is S4-3_21 (student with index number 21 in the third class in secondary 4)
‘S1_Class’	Nominal	The class that the student was in during Secondary 1
‘S2_Class’	Nominal	The class that the student was in during Secondary 2
‘S3_Class’	Nominal	The class that the student was in during Secondary 3
‘S4_Class’	Nominal	The class that the student was in during Secondary 4
‘OLVL_Batch’	Nominal	The year in which the student took the ‘O’ Levels
‘2nd_Lang’	Nominal	The second language of the student (Chinese. Malay etc.)

‘MT_Allocated’	Nominal	The mother tongue language that the student is undertaking (Chinese, Higher Chinese, Malay etc.)
‘Year_Transferred’	Continuous	The year that the student was transferred to this secondary school. Blank implies that the student was not a transfer student
‘Subject_Combination’	Nominal	The subject combination that the student took (Triple Science, Double Science, 1 Pure 1 Combined, Combined Science)
‘CCA_Name_1’	Nominal	The name of the a CCA the student is involved in (it is compulsory that students in this secondary school be involved in at least one CCA)
‘CCA_Name_2’	Nominal	The name of the second CCA the student is in (can be blank)
‘CCA_Name_3’	Nominal	The name of the third CCA the student is in (can be blank)
‘CCA_L’	Continuous	The CCA points awarded to the student in the area of Leadership
‘CCA_E’	Continuous	The CCA points awarded to the student in the area of Enrichment
‘CCA_A’	Continuous	The CCA points awarded to the student in the area of Achievement
‘CCA_P’	Continuous	The CCA points awarded to the student in the area of Participation

‘CCA_S’	Continuous	The CCA points awarded to the student in the area of Service
‘CCA_Total_Points’	Continuous	The sum total of the CCA points that the student attained
‘CCA_Grade’	Continuous	The corresponding grade that the student attained. The lower the better
‘P6_PSLE_L1’	Nominal	The letter grade awarded to the student for the score of his first language
‘P6_PSLE_L2’	Nominal	The letter grade awarded to the student for the score of his second language
‘P6_PSLE_SC’	Nominal	The letter grade awarded to the student for the score of his science subject
‘P6_PSLE_MA’	Nominal	The letter grade awarded to the student for the score of his mathematics subject
‘P6_PSLE_AGG’	Continuous	The aggregate score of the student’s overall Primary School Leaving Examination (max score is 300)
‘S1_CA1_CL’	Continuous	The score for the student’s Chinese Language during CA1 (Continual Assessment 1) when the student was in Secondary 1
‘S1_CA1_D&T’	Continuous	The score for the student’s Design & Technology during CA1 (Continual Assessment 1) when the student was in Secondary 1

‘S1_CA1_EL1’	Continuous	The score for the student’s English Language during CA1 (Continual Assessment 1) when the student was in Secondary 1
‘S1_CA1_LIT(E)’	Continuous	The score for the student’s Literature during CA1 (Continual Assessment 1) when the student was in Secondary 1
‘S1_CA1_MATHS’	Continuous	The score for the student’s Mathematics during CA1 (Continual Assessment 1) when the student was in Secondary 1
‘S1_CA1_ML’	Continuous	The score for the student’s Malay Language during CA1 (Continual Assessment 1) when the student was in Secondary 1
‘S1_CA1_GEOG’	Continuous	The score for the student’s Geography during CA1 (Continual Assessment 1) when the student was in Secondary 1
‘S1_CA1_SCI(S/E)’	Continuous	The score for the student’s Science during CA1 (Continual Assessment 1) when the student was in Secondary 1
‘S1_CA1_HIST’	Continuous	The score for the student’s History during CA1 (Continual Assessment 1) when the student was in Secondary 1
‘S1_CA1_ART’	Continuous	The score for the student’s Art during CA1 (Continual Assessment 1) when the student was in Secondary 1

<table>
<tr><td>‘S1_CA1_OVERALL’</td><td>Continuous</td><td>The overall score of the student (max score = number of subjects undertaken by student * 100)</td></tr>
<tr><td>‘S1_CA1_AVERAGE’</td><td>Continuous</td><td>The average score of the student (overall score divided by the number of subjects undertaken by the student)</td></tr>
<tr><td colspan="3">Note: The above data for Secondary 1 CA1 will follow in a similar pattern for the rest of the four years of the student’s studies. The order will be S1_CA1, S1_SA1 (Semester Assessment), S1_CA2, S1_SA2, S1_OVL (the overall score of the student’s performance in the entire academic year, with various percentages from the CA1, SA1, CA2 & SA2).
This will repeat from S1 to S4. In Secondary 3 (S3) and Secondary 4 (S4), there will be more subjects available to the student and a blank would indicate that the student did not undertake that subject during that exam.
Finally, at the end of S4, there will be another set of similar pattern of data for the student’s ‘O’ Levels examinations. It will follow the format of S4_OLVL_<subjectName>. So an example would be S4_OLVL_GEOG for the score of Geography taken during ‘O’ Levels.</td></tr>
<tr><td>‘S4_OLVL_L1R4’</td><td>Continuous</td><td>The ‘O’ Level score attained by the student by taking the number in the letter grade of a language subject and FOUR other relevant subject (humanities, sciences, mathematics). The lower the better</td></tr>
</table>

‘S4_OLVL_L1R5’	Continuous	The ‘O’ Level score attained by the student by taking the number in the letter grade of a language subject and FIVE other relevant subject (humanities, sciences, mathematics). The lower the better
Note: After the ‘S4_OLVL_L1R5’, there is the set of standardized data which follows the similar format as of the S1 to S4 results.		