

Using Multiple Linear Regression to Analyze Factors Affecting Differences in Results for the Programme for International Student Assessment (PISA) Global Education Survey across Schools in Singapore

Chermain Ang; Gareth Shaun Ng Wei Long; Ong Qinghua Jeremy,
Singapore Management University

ABSTRACT

The Organisation for Economic Co-operation and Development's (OECD) Programme for International Student Assessment (PISA) global education survey is a triennial international survey that aims to evaluate education systems worldwide by testing the skills and knowledge of 15-year-old students in Mathematics, Reading, and Science. The survey has become increasingly influential on politicians who see their countries and their policies being measured against these global school league tables. At the same time, there have been many discussions and debates in Singapore regarding the issues surrounding the ever-evolving education system. In this paper, we used multiple linear regression to build an explanatory model to find out what affects schools' overall scores (Reading, Mathematics and Science), as well as science-specific scores. The explanatory variables are derived from questions in the school questionnaire distributed to school principals. We also seek to find out if it is possible for all schools to start on an equal footing, and the potential steps stakeholders such as the Ministry of Education (MOE) can potentially take in order to even the playing field, as well as improve schools' performance in general. Based on our results, we are proposing three key recommendations to schools and the education ministry – to increase training and development for teachers, fine-tune the selection process for hiring teachers, and to enhance parents' involvement in school activities through meaningful engagement.

INTRODUCTION

In the recently released 2015 Programme for International Student Assessment (PISA) results, Singapore students topped the test in all three areas – Mathematics, Reading and Science. It is Singapore's best performance in the international assessment thus far. This paper aims to identify factors affecting Singapore schools' performance using two key metrics, namely each school's mean overall score (covers all booklets – Reading, Mathematics and Science questions included), as well as each school's mean science score (Science questions only). Science scores were analyzed in addition to overall scores since all booklets contained Science questions, whereas not all booklets had Reading and Mathematics questions. In other words, not all schools were evaluated for Reading and Mathematics, and hence it was possible to derive only the mean school science score for every school that took part in this survey.

There has been a lot of discussion among parents, educators and the general public on the issue of disparity across schools with regard to academics. Singapore's Minister of Education, Mr Heng Swee Keat initiated a slogan for Singapore schools, "every school a good school" in 2013. Furthermore, OECD education director Andreas Schleicher shared in a BBC article that "Singapore managed to achieve excellence without wide differences between children from wealthy and disadvantaged families." However, the public sentiment is that all students do not start on an equal footing; more support can and should be given to students from less privileged backgrounds. Therefore, we seek to explore the factors contributing to the differences in overall scores and science scores across all schools.

The rest of the paper is organized as follows. Following the introduction, we review the literatures related to our study. This is followed by an overview of our methodology and our data preparation steps. In the next section, the various methods and techniques used to narrow down the explanatory variables will be discussed, and following that will be the stepwise multiple linear regression process and our insights from the model. Lastly, the paper concludes by highlighting the key recommendations to schools and the education ministry.

LITERATURE REVIEW

There has been numerous research done across the world using the PISA results, which is released once every three years. Most of the research are done at an international level, and while there are country-specific research, there are minimal research done on Singapore's results. Therefore, we are interested in analyzing Singapore's results to find out if there are similarities and differences.

There are multiple findings stating that a student's performance is generally better when their socioeconomic status is higher [6], and socioeconomically advantaged students tend to get better scores as compared to their disadvantaged peers regardless of countries and economies [2]. Naturally, drawing it back to the comparison schools' performance,

it can be hypothesized that schools with greater percentage of disadvantaged students from a socioeconomic perspective tend to perform more poorly overall.

We decided to use multiple linear regression as the main technique to determine the correlations between the mean school overall or science score and the questions in the school questionnaire filled in by the principal or relevant school personnel. Past research has also used the regression model to analyze and even predict how well students will do for a specific subject such as Mathematics [5]. Rather than a predictive model, we intend to create an explanatory model to analyze variables affecting schools' performance.

METHODOLOGY



Figure 1. Flow diagram illustrating analytical process

Figure 1 illustrates the analytical process used for this paper. After data preparation, we proceeded with the data analysis using several analytical techniques – since the dataset contains both continuous and categorical explanatory variables, there is a need to separate the two types of explanatory variables during the initial feature selection, prior to conducting the stepwise regression model. For the continuous explanatory variables, we used the standard least squares regression method to remove correlated variables. For the categorical explanatory variables, we will be using decision tree for feature selection. Next, the team conducted multiple linear regression to identify and analyze the factors that affect the scores of the schools, using the observations from the data analysis segment to provide key insights and recommendations.

A regression model is a mathematical model that explains and predicts a continuous response variable. For our analysis, a regression model will be developed to explain why certain schools score better than others. Multiple linear regression is the key technique selected to derive our insights due to its flexibility in allowing us to use both continuous and categorical variables. In this case, the explanatory variables are derived from the questions posted to the school, and the response variables are the schools' mean overall score and schools' mean science score, which will be analyzed separately.

The 2015 PISA data was released last December 6 2016 and it will be used for our analysis. More specifically, we did this analysis based on the final data output after standardization of scores across all booklets, with reference to Paper XXX-2017.

The above-mentioned data analysis will be carried out using JMP Pro 13, which provides all the techniques we require. Its in-memory processing features also allowed us to run iterations of the various analyses multiple times at an efficient speed when necessary.

DATA PREPARATION

STEP 1: SORTING EXPLANATORY VARIABLES BY TYPE

Using the codebook provided by OECD, the team sorted the questions from the school questionnaire into continuous, ordinal or nominal variables by observing the question types. As shown below, Figures 2, 3 and 4 illustrate examples of a continuous, ordinal and nominal variable respectively.

SC018Q01TA0 Teachers in TOTAL: Full-time	NUM	5.0	138	0 - 1327
--	-----	-----	-----	----------

Figure 2. Example of continuous explanatory variable

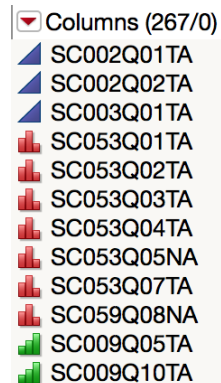
SC009Q08TA Frequency of <the last academic year>. I pay attention to disruptive behaviour in classrooms.	1	Did not occur
	2	1-2 times during the year
	3	3-4 times during the year
	4	Once a month
	5	Once a week
	6	More than once a week

Figure 3. Example of ordinal explanatory variable

SC059Q06NA We have enough laboratory material that all courses can regularly use it.	1	Yes
	2	No

Figure 4. Example of nominal explanatory variable

As our analysis will be done using JMP Pro 13, those changes were on the data table through the software as well, as shown in Figure 5 below.



A screenshot of the JMP Pro 13 software interface showing a list of columns. The list is titled 'Columns (267/0)'. The variables are listed with corresponding icons: blue triangles for continuous variables and red or green bar charts for nominal and ordinal variables. The variables shown are SC002Q01TA, SC002Q02TA, SC003Q01TA, SC053Q01TA, SC053Q02TA, SC053Q03TA, SC053Q04TA, SC053Q05NA, SC053Q07TA, SC059Q08NA, SC009Q05TA, and SC009Q10TA.

Columns (267/0)
SC002Q01TA
SC002Q02TA
SC003Q01TA
SC053Q01TA
SC053Q02TA
SC053Q03TA
SC053Q04TA
SC053Q05NA
SC053Q07TA
SC059Q08NA
SC009Q05TA
SC009Q10TA

Figure 5. Examples of different types explanatory variable

STEP 2: EXCLUDING VARIABLES WITH MISSING VALUES

The team also removed both response and explanatory variables with too many missing values. For the response variable, school ID 29 was removed. For the explanatory variables, we set an arbitrary threshold – no more than 20%, or 35.4 out of 177 data points should be missing. Based on the threshold, we excluded “SC014Q01NA”.

REMOVING CORRELATED CONTINUOUS EXPLANATORY VARIABLES

For the remaining continuous explanatory variables, we conducted standard least squares regression to identify and

exclude correlated variables through observing the correlation of estimates (Figure 6), ensuring that they do not exceed a threshold of +/- 0.7.

Multivariate							
Correlations							
	SC002Q01TA	SC002Q02TA	SC003Q01TA	SC004Q01TA	SC004Q02TA	SC004Q03TA	SC004Q04NA
SC002Q01TA	1.0000	-0.2327	0.0286	0.2784	-0.0267	-0.0268	-0.1190
SC002Q02TA	-0.2327	1.0000	0.1203	0.2068	0.2119	0.2084	0.1523
SC003Q01TA	0.0286	0.1203	1.0000	0.3660	0.2635	0.2746	0.2101
SC004Q01TA	0.2784	0.2068	0.3660	1.0000	0.3320	0.3311	0.2281
SC004Q02TA	-0.0267	0.2119	0.2635	0.3320	1.0000	0.9972	0.8678
SC004Q03TA	-0.0268	0.2084	0.2746	0.3311	0.9972	1.0000	0.8630
SC004Q04NA	-0.1190	0.1523	0.2101	0.2281	0.8678	0.8630	1.0000
SC004Q05NA	0.0261	-0.0735	-0.3623	-0.2836	-0.2188	-0.2208	-0.1549
SC004Q06NA	0.3575	0.1784	-0.0648	0.2362	0.2830	0.2849	0.1796

Figure 6. Table showing correlation of estimates of sampled variables from the first iteration of standard least square regression of overall scores given all continuous variables

The variables were removed conservatively, as we aim to retain as many variables as possible, in order to avoid missing out on variables that might have a huge effect on the response variable. Three iterations of standard least squares regression were done to ensure that no remaining variables were correlated. This was further confirmed by checking the Variance Inflation Factors (VIF), as shown in Figure 7 below. VIF is useful in determining multicollinearity within variables. While there are no formal criteria with regard to an acceptable level of VIF, a common recommendation is a value of ten; and a clear signal of multicollinearity is when VIF is greater than eight. However, it is also important to pay attention to variables that have a VIF of five or more. In this case, as seen in Figure 7, the final set of selected variables have VIF values of less than five, indicating that multicollinearity does not exist in the final iteration of our standard least squares regression model.

Response Mean(Standardized Scoring)					
Parameter Estimates					
Term	Estimate	Std Error	t Ratio	Prob> t	VIF
Intercept	0.3916704	0.071393	5.49	<.0001*	.
SC004Q01TA	7.2371e-5	0.000101	0.72	0.4748	1.7009541
SC004Q02TA	-2.17e-5	0.000154	-0.14	0.8880	3.58459
SC004Q04NA	0.0001028	0.000132	0.78	0.4404	3.0475662
SC004Q05NA	-0.002128	0.001008	-2.11	0.0380*	1.5680619
SC004Q06NA	0.0001241	0.000362	0.34	0.7325	2.0557817
SC004Q07NA	6.2143e-5	0.000106	0.59	0.5584	1.2586255
SC016Q01TA	0.0001582	0.000547	0.29	0.7732	1.7319271
SC018Q05NA01	-4.167e-5	0.000493	-0.08	0.9329	2.3554889
SC018Q05NA02	0.0059182	0.002708	2.19	0.0320*	1.3912684
SC018Q06NA01	0.0015814	0.000842	1.88	0.0644	3.2680948
SC018Q06NA02	-0.009765	0.007533	-1.30	0.1988	1.7459349
SC018Q07NA02	0.1366489	0.047398	2.88	0.0051*	1.3756713
SC019Q03NA01	0.0037439	0.001626	2.30	0.0241*	3.3178211
SC019Q03NA02	-0.001898	0.009111	-0.21	0.8355	1.4323192
SC048Q01NA	0.0004739	0.000281	1.69	0.0956	1.2995682
SC048Q02NA	-0.000291	0.001847	-0.16	0.8752	1.5322895
SC048Q03NA	-0.004281	0.000873	-4.90	<.0001*	1.7201939
SC064Q01TA	0.0001316	0.000306	0.43	0.6684	1.1726238
SC064Q02TA	-0.000193	0.000316	-0.61	0.5423	1.248144
SC064Q03TA	-0.000432	0.000907	-0.48	0.6347	1.4140444
SC064Q04NA	0.0012861	0.000794	1.62	0.1096	1.4411688
SC025Q01NA	2.5281e-5	0.000327	0.08	0.9386	1.2461476

Figure 7. Table showing Variance Inflation Factor (VIF) of variables from the final iteration of standard least square regression of overall scores given selected continuous variables

After three iterations of standard least squares regression for both response variables, there was a final number of 22 continuous explanatory variables for schools' mean overall scores, and 21 continuous explanatory variables for schools' mean science scores. These variables will be used for the final step, stepwise regression.

FEATURE SELECTION OF CATEGORICAL EXPLANATORY VARIABLES USING DECISION TREE

Due to the excessive number of categorical explanatory variables, instead of including all of them in the stepwise multiple linear regression model, we used decision tree to conduct feature selection, whereby the variables which are important and affect the response variables will be selected for stepwise regression. The number of splits is determined by ensuring that for each split conducted, the R-square value continues to rise and does not reach a plateau by observing the split history graph as seen in Figure 8 below. In our case, it reached saturation prior to the graph reaching a plateau.

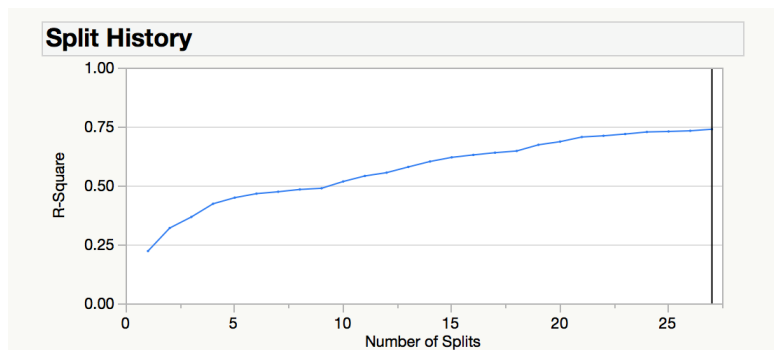


Figure 8. Graph showing number of splits against R-square for decision tree (Overall Scores)

The selection of variables is determined by the logworth of the variable, whereby all variables with positive logworth (greater than zero) will be selected, as seen in Figures 9 and 10 below.

Column Contributions				
Term	Number of Splits	SS		Portion
SC012Q06TA	1	0.68262638		0.3032
SC053Q05NA	1	0.296485		0.1317
SC035Q10TB	1	0.16964001		0.0753
SC061Q01TA	1	0.14294961		0.0635
SC063Q04NA	2	0.10781184		0.0479
SC010Q06TC	2	0.10367029		0.0460
SC034Q04TA	2	0.08628422		0.0383
SC009Q05TA	1	0.07864303		0.0349
SC053Q07TA	1	0.07335942		0.0326
SC012Q02TA	1	0.07142401		0.0317
SC010Q04TE	1	0.07023804		0.0312
SC010Q12TE	1	0.05318189		0.0236
SC010Q02TE	1	0.05165119		0.0229
SC059Q08NA	1	0.04238636		0.0188
SC010Q01TC	1	0.04081201		0.0181
SC032Q04TA	1	0.03176841		0.0141
SC010Q02TA	1	0.02975941		0.0132
SC009Q10TA	2	0.0287976		0.0128
SC037Q09TA	1	0.02434131		0.0108
SC017Q01NA	1	0.02246203		0.0100
SC010Q09TE	1	0.02199972		0.0098
SC035Q11NB	1	0.0147574		0.0066
SC010Q10TB	1	0.006368		0.0028

Figure 9. Table showing categorical variables with positive logworth values (Overall Scores)

Column Contributions			
Term	Number of Splits	SS	Portion
SC012Q06TA	1	0.65546105	0.2985
SC053Q05NA	1	0.32151442	0.1464
SC061Q01TA	1	0.13411749	0.0611
SC035Q10TB	1	0.13324499	0.0607
SC010Q06TC	1	0.08926487	0.0407
SC010Q04TE	1	0.07699906	0.0351
SC012Q02TA	1	0.06949737	0.0316
SC009Q05TA	1	0.06889302	0.0314
SC042Q01TA	1	0.06750794	0.0307
SC053Q07TA	1	0.06653679	0.0303
SC027Q04NA	1	0.06652902	0.0303
SC059Q04NA	1	0.04908048	0.0224
SC010Q02TE	1	0.04550118	0.0207
SC059Q08NA	1	0.0419185	0.0191
SC034Q04TA	2	0.03829725	0.0174
SC035Q07TB	1	0.03621376	0.0165
SC010Q12TE	1	0.03554109	0.0162
SC010Q02TA	1	0.02846548	0.0130
SC010Q01TC	1	0.0281753	0.0128
SC037Q09TA	1	0.026535	0.0121
SC009Q10TA	1	0.02341966	0.0107
SC010Q11TB	1	0.02177892	0.0099
SC009Q02TA	1	0.02040602	0.0093
SC017Q02NA	1	0.01879477	0.0086
SC035Q11NB	1	0.01775729	0.0081
SC061Q08TA	1	0.01327298	0.0060
SC059Q02NA	1	0.00119072	0.0005

Figure 10. Table showing categorical variables with positive logworth values (Science Scores)

STEPWISE MULTIPLE LINEAR REGRESSION MODEL

SELECTION OF DIRECTION FOR STEPWISE REGRESSION

After the above feature selection processes, 22 continuous variables and 23 categorical variables were used for the regression model for the mean school overall scores, while 21 continuous variables and 27 categorical variables were used for the mean school science scores. Backward, forward and mixed stepwise regression models were generated, where a selection criteria for a variable to enter or leave was if they had a p-value of less than 0.05 for both schools' mean overall score and schools' mean science score.

Fit Group	
Response Mean(Standardized Scoring)	
Summary of Fit	
RSquare	0.740088
RSquare Adj	0.705586
Root Mean Square Error	0.071246
Mean of Response	0.521251
Observations (or Sum Wgts)	129

Figure 11. Table showing backward stepwise regression model's Summary of Fit (Overall Scores)

Fit Group	
Response Mean(Science %)	
Summary of Fit	
RSquare	0.733729
RSquare Adj	0.690935
Root Mean Square Error	0.071291
Mean of Response	0.513482
Observations (or Sum Wgts)	131

Figure 12. Table showing backward stepwise regression model's Summary of Fit (Science Scores)

Upon comparison of the three methods, backward stepwise regression results in the highest adjusted R-square for both mean school overall scores (adjusted R-square of 0.7056) and mean school science scores (adjusted R-square of 0.6909), as seen in Figure 11 and 12 above. In other words, the set of explanatory variables as seen in Figure 13 can account for 70.56% of the variation in the mean school overall scores, and the explanatory variables found in Figure 14 explains 69.09% of the variation in the mean school science scores. Given that the variables derived from the backward stepwise regression model allows us to best explain the variation in the schools' performance, the results from backward stepwise regression will be used for the analysis.

INSIGHTS FROM STEPWISE REGRESSION MODEL

The full results from the backward stepwise regression model can be found in Figures 13 and 14 for both schools' mean overall score and schools' mean science score respectively. If the variable was nominal, and the term is shown as, for instance "SC053Q07TA[1]", the number in the square brackets represent the response option involved. For ordinal variables, the term displayed in a format such as "SC037Q09TA{1-3&2}", the involved response options will be shown in the braces. In the example "SC037Q09TA{1-3&2}", the response option 1 (excluding response options 2 and 3) of the variable results in the differences in scores.

Variables Affecting Overall Scores

Fit Group				
Response Mean(Standardized Scoring)				
Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	0.5241833	0.029769	17.61	<.0001*
SC018Q07NA02	0.0782989	0.038733	2.02	0.0456*
SC053Q07TA[1]	0.0322451	0.009551	3.38	0.0010*
SC053Q05NA[1]	0.0233774	0.007737	3.02	0.0031*
SC063Q04NA[1]	0.0135695	0.006594	2.06	0.0419*
SC009Q05TA{3&4-5&6}	0.0122029	0.007147	1.71	0.0905
SC034Q04TA{2&1-4&3&5}	0.0058602	0.007109	0.82	0.4115
SC019Q03NA01	0.0044681	0.000794	5.63	<.0001*
SC035Q11NB[1]	-0.001072	0.006711	-0.16	0.8733
SC004Q05NA	-0.0024	0.000721	-3.33	0.0012*
SC048Q03NA	-0.004258	0.000614	-6.93	<.0001*
SC034Q04TA{2-1}	-0.01131	0.009707	-1.17	0.2464
SC009Q05TA{2-3&4&5&6}	-0.017561	0.01426	-1.23	0.2207
SC010Q01TC[0]	-0.018819	0.008205	-2.29	0.0237*
SC009Q05TA{3-4}	-0.020274	0.008714	-2.33	0.0218*
SC037Q09TA{1-3&2}	-0.025456	0.015931	-1.60	0.1129

Figure 13. Table showing variables from backward stepwise regression model sorted by parameter estimates in descending order (Overall Scores)

Term	Question	Response Options	Estimate (Overall)
SC063Q04NA[1]	School includes parents in school decisions.	1 Yes 2 No	0.0135695
SC018Q07NA02	Teachers with an <ISCED Level 6> qualification: Part-time	(continuous variable)	0.0782989

Table 1. Table showing selected variables with significance levels of less than 0.05 for overall scores

Parents involvement in school decisions (SC063Q04NA)

It is recommended that schools include parents in their decision-making process for school-related issues, as schools that have chosen to include parents have fared better at the PISA results.

This is in line with recent trends where schools aim to engage parents beyond the "superficial" purposes such as fundraising or attending events. One potential reason for this variable to be significant to the schools' mean overall scores is that parents feel more ownership when they get to participate in school decisions, encouraging them to contribute their valuable knowledge, skills and viewpoints.

Education level of part-time teachers (SC018Q07NA02)

Another interesting insight is that having a greater number of part-time teachers with a degree from a second stage of tertiary education, such as masters or doctoral degree, results in better overall scores. This supports the finding below as seen in Table 3, whereby it is encouraged for schools to hire teachers with tertiary education qualifications.

Variables Affecting Science Scores

Fit Group				
Response Mean(Science %)				
Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	0.5779303	0.040173	14.39	<.0001*
SC053Q07TA[1]	0.0300571	0.009734	3.09	0.0025*
SC053Q05NA[1]	0.0269779	0.007604	3.55	0.0006*
SC009Q10TA{2-3&4&5&6}	0.0240565	0.02022	1.19	0.2367
SC009Q05TA{3&4-5&6}	0.011446	0.007746	1.48	0.1423
SC019Q03NA01	0.004913	0.001129	4.35	<.0001*
SC034Q04TA{2&1-4&3}	0.0036358	0.008263	0.44	0.6608
SC035Q07TB[1]	0.0026934	0.009043	0.30	0.7664
SC064Q04NA	0.0012142	0.000683	1.78	0.0783
SC025Q02NA	0.0004049	0.000209	1.93	0.0557
SC018Q05NA01	-0.00047	0.000393	-1.20	0.2345
SC034Q04TA{2&1&4&3-5}	-0.001012	0.007837	-0.13	0.8975
SC064Q03TA	-0.001452	0.0008	-1.82	0.0721
SC004Q05NA	-0.002504	0.000741	-3.38	0.0010*
SC048Q03NA	-0.004568	0.000633	-7.22	<.0001*
SC009Q05TA{2-3&4&5&6}	-0.009155	0.015953	-0.57	0.5672
SC009Q05TA{3-4}	-0.012872	0.008621	-1.49	0.1382
SC010Q01TC[0]	-0.023162	0.008623	-2.69	0.0083*
SC034Q04TA{2-1}	-0.02366	0.010456	-2.26	0.0256*

Figure 14. Table showing variables from backward stepwise regression model sorted by parameter estimates in descending order (Science Scores)

Term	Question	Response Options	Estimate (Overall)
SC025Q02NA	Teaching staff in your school has attended a programme of profess dev? Science teaching staff	(continuous variable)	0.0004049
SC064Q04NA	<the last academic year>, what proport. of parents part. school-related activities? Volun\phys, or extra-curricular act	(continuous variable)	0.0012142
SC009Q10TA{2-3&4&5&6}	Frequency of <the last academic year>. I engage teachers to help build a school culture of continuous improvement.	1 Did not occur 2 1-2 times during the year 3 3-4 times during the year 4 Once a month 5 Once a week 6 More than once a week	0.0240565

Table 2. Table showing selected variables with significance levels of less than 0.05 for science scores

Participation in professional development programmes for teachers (SC025Q02NA)

It is comforting to note that having greater number of science teachers attending professional development programmes contribute to better school scores, as it shows that these programmes are effective in preparing the teachers to become better educators, allowing the students to learn more effectively.

Proportion of parents' participation in school-related activities (SC064Q04NA)

Similar to the previous finding for overall scores where parents' participation contributes to better school results (overall scores), the greater the proportion of parents participating in school-related activities such as volunteering, the better the school's performance in science.

Frequency of principal's engagement with teachers to create a school culture of continuous improvement (SC009Q10TA)

Intriguingly, there is an ideal frequency for principals to engage their teachers to create a school culture of continuous improvement, which is "1-2 times during the year". This shows that it is important for principals or leaders to remind teachers of the need to continuously improve, and that status quo is never good enough. However, at the same time,

it is critical to not do it too often, as it may potentially divert too much time and effort from other important matters such as time spent on the curriculum or teaching methods.

Variables Affecting Both Overall Scores And Science Scores

There are 11 variables affecting both the schools' mean overall score and schools' mean science score relatively significantly, and five of them are displayed below in Table 3.

Term	Question	Response Options	Estimate (Overall)	Estimate (Science)
SC010Q01TC[0]	Selecting teachers for hire: <School governing board>	0 Not checked 1 Checked	-0.018819	-0.023162
SC048Q03NA	Est. percent. <national modal grade for 15-year-olds>. Students from socioeconomic disadvantaged homes	(continuous variable)	-0.004258	-0.004568
SC019Q03NA01	<School science> teachers\<ISCED Level 5A or higher> qualification <with a major> in <school science>: Full-time	(continuous variable)	0.0044681	0.004913
SC053Q05NA[1]	<This academic year>,follow. activities\school offers<national modal grade for 15-year-olds>? Science club	1 Yes 2 No	0.0233774	0.0240565
SC053Q07TA[1]	<This academic year>,follow. activities\school offers<national modal grade for 15-year-olds>? Chess club	1 Yes 2 No	0.0322451	0.0269779

Table 3. Table showing selected variables with significance levels of less than 0.05 for both overall scores and science scores

Significance of extra-curricular activities (SC053Q05NA & SC053Q07TA)

As illustrated in Table 3, schools that offer extra-curricular activities, specifically Science Club and Chess Club tend to do better. These variables are also two of the variables with the highest absolute parameter estimate values, indicating that they have a relatively more significant effect on the mean schools' scores.

Therefore, the presence of these extra-curricular activities clubs is a good determinant of the school's capabilities, potentially due to the fact that these clubs enrich the students' learning and growth through activities that engage their minds effectively.

Percentage of students from socioeconomic disadvantaged homes (SC048Q03NA)

Schools with a higher percentage of students from socioeconomic disadvantaged homes tend to do less well in the PISA survey. This is in line with past research, which has shown that socio-economic status does affect a student's performance, whereby "home background makes a substantial contribution to student differences".

This further illustrates the need for relevant stakeholders such as the government, more specifically the Ministry of Education, to ensure that students from socioeconomic disadvantaged homes are given sufficient support to start on an equal footing, and to be given the chance to reach their full potential despite coming from a less privileged background. In the context of schools, this can be done by identifying schools with higher percentage of socioeconomic disadvantage families, and providing more subsidies or grants for free tuition or enrichment courses. This is especially the case in Singapore, where more than 60% of parents of secondary school children, the target age group for this survey, send their children for tuition.

Role of school governing board in selection of teachers for hire

Interestingly, the school governing board should ideally play a part in selecting teachers for hire, since schools that did not include the school governing board in the selection process tend to do worse. This may be due to the lower level of structure or lower standards in the selection process for hiring teachers if the school governing board was not involved. Another potential reason is the lack of experience within the hiring panel if the school governing board were to be left out of the process.

Education level of full-time science teachers (SC019Q03NA01)

Schools with a greater number of full-time school science teachers with minimally a bachelor's degree tend to do better. As expected, this variable has a greater impact on the schools' mean science scores compared to the overall scores.

This implies that education level of the teachers do affect their students' performance, likely due to the way they teach or conduct lessons, given that the content of the curriculum is held constant. Therefore, schools that wish to see better academic results can consider investing in hiring more teachers with a bachelor's degree.

CONCLUSION

Given that one of the contributing variables affecting school performance is the percentage of students from socioeconomic disadvantaged backgrounds, it is a telltale sign that there is indeed a difference across schools with regard to their starting ground. Therefore, to ensure that all schools can provide the same support to their students, the Ministry of Education (MOE), as well as the schools themselves, can consider our recommendations in the following three broad areas:

1. Training and Development for teachers
2. Fine-tuning the selection process for hiring teachers
3. Increasing parents' involvement through meaningful engagement

For training and development, the school can focus on professional development courses aimed at improving the overall quality of teaching across all teaching staff. Schools should not have to decide on budget allocation between supporting students with less privileged background and training programmes for teachers. Ideally, MOE should aim to provide more grants to schools with a greater percentage of less privileged students, with the specific purpose of ensuring that the students from socioeconomic disadvantaged backgrounds get the support they need, be it in terms of having a wholesome meal at school, or attending enrichment courses, which has become a norm in Singapore.

With regard to the selection process for hiring teachers, MOE can consider allocating the talent pool of teachers with tertiary education equally across all schools. Furthermore, from the results, it can be seen that the school governing body should play a role in the selection process of teachers as well.

Finally, parents' involvement in school activities should be encouraged as it increases the parents' sense of ownership in their children's education journey, allowing them to feel more invested and hence dedicate more effort and time to guiding and educating their child academically.

ACKNOWLEDGEMENTS

We would like to show our appreciation to Prof Kam Tin Seong (Associate Professor of Information Systems; Senior Advisor, SIS) for guiding us throughout this process of data preparation, analysis and insights generation.

REFERENCES

- [1] Classifying educational programmes: manual for ISCED-97 implementation in OECD countries. (1999). Paris: Organisation for Economic Co-operation and Development.
- [2] Grade expectations: how marks and education policies shape students' ambitions. (2012). Retrieved April 10, 2017, from https://play.google.com/store/books/details?id=YdM3-8fwcEwC&rdid=book-YdM3-8fwcEwC&rdot=1&source=gbs_vpt_read&pcampaignid=books_booksearch_viewport
- [3] Learning for tomorrow's world: first results from PISA 2003. (2004). Paris: Organisation for Economic Co-operation and Development.
- [4] School factors related to quality and equity: results from PISA 2000. (2005). Paris: Organisation for Economic Co-operation and Development.
- [5] Shakil, M. (n.d.). To Predict the Student's Final Grade in a Mathematics Class. Retrieved April 15, 2017, from <http://www.shsu.edu/~wxb001/documents/Amultipleregressionmodelpaper.pdf>
- [6] Sweet, R., Nissinen, K., & Vuorinen, R. (2014). An analysis of the career development items in PISA 2012 and of their relationship to the characteristics of countries, schools, students and families. Retrieved April 19, 2017, from <http://www.elgpn.eu/publications/browse-by-language/english/elgpn-research-paper-no.-1-pisa/>

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.