Paper ###-2017

Using Partition Models to Identify Key Differences Between Top Performing and Poor Performing Students 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

With Singapore students topping the PISA test in 2015, our project sponsor is interested to find out what factors contribute to the success of top performing students, and what are the characteristics of the poorer performing students. We performed standardized scoring on the student cognitive questionnaire data extracted from PISA's online database, and combined it with the student questionnaire data containing students' response to their family and personal background. With the combined data, we aim to build an explanatory model to identify the key factors influencing student's performance in the PISA test. Partition Models in JMP Pro 13, namely Regression Tree, Boosted Tree, and Bootstrap Forest are considered for our analysis. The Boosted Tree model is found to be the best model, with a high RSquared value and minimal difference between the RMSE values for the Training and Validation sets. We share our findings on the insights gleaned from our using the Partition models, as well as the key factors identified, which serves to explain the performance of Singapore students in the 2015 PISA test.

INTRODUCTION

Being products of the Singapore education system ourselves, the team would like to assess the effectiveness of Singapore's education system in student development, following the comment made by OECD's education director, Andreas Schleicher, that "Singapore managed to achieve excellence without wide differences between children from wealthy and disadvantaged families." In a BBC article. One of the main motivation behind this paper is to verify the claims made, as well as to identify factors schools and parents can work on to improve student's performance.

In the 2015 edition of the Programme for International Student Assessment (PISA) global education survey conducted by the Organization for Economic Co-operation and Development's (OECD), Singapore students achieved the best results yet, outperforming the world in all three subjects – Science, Mathematics, and Reading. Held every three years, the survey evaluates a country's school system with regards to its quality, equity, and efficiency. The raw data used for this paper is obtained from the PISA 2015 Database online. The output of the Standardized Scoring of the Cognitive Questionnaire dataset performed in Paper XXX-2017 will also be used in this paper. Having achieved such stellar performance on the world stage, this paper aims to uncover the ingredients of success for Singapore's top performers, and what makes them different from the poor performers for each subject as well as their overall performance in the test, through recursive partitioning methods such as Decision Tree, Bootstrap Forest, and Boosted Tree.

The paper will continue with literature reviews of other similar works, followed by an overview of the methodology, the data preparation steps, model planning considerations, and the discussion of results from the Boosted Tree model. Lastly, key findings on the differing characteristics between the top and poorer performing students are highlighted in the concluding paragraph.

LITERATURE REVIEW

Past works using Decision Tree models on Education datasets focuses more on using the models as a predictive tool instead of an explanatory tool. In the work by Priyanka and team [1], they study various Decision Tree algorithms to evaluate the best one for classifying educational data. Harwait and Amby [2], used Decision Tree models to determine the characteristics to consider for admission in building a student admission selection model. Mashael and Muna [3] applied Decision Tress to predict student's final GPA based on student's grades in previous courses. Quadril and Kalyankar [4] adopted Decision Tree techniques to predict and identify students who are more likely to drop out of school based on past records, and using the predictions as a guide for schools and educators to identify appropriate strategies to prevent the student from dropping out of school. These works offer interesting insights on how Decision Tree models can be used to efficiently on education datasets in model construction for analyzing, explaining, or predicting the data available.

In Partition models, the data is recursively partitioned according to the relationship between the dependent and independent variables to form a decision tree. The benefits of using such models is that the results are easy to interpret, it is able to handle a large amount of data, and does not require a prior good model for us to explore the relationships within the dataset.

Another type of partition modelling is that of a Bootstrap Forest (Random decision forests), first created by Tin Kam Ho. In this model, many decision trees are built, and subsequently combined to form a more powerful model. It averages the outcome of all the decision trees built to arrive at the final model output.

Boosted Tree is another partition model used in our analysis. It is a process of fitting many small decision trees sequentially (layer by layer), to build a large decision tree. As the tree fits layer by layer, it corrects the poor fitting of data from the previous layers, by fitting according to the residuals of previous layers. The final output is the sum of the residuals across all layers.

METHODOLOGY

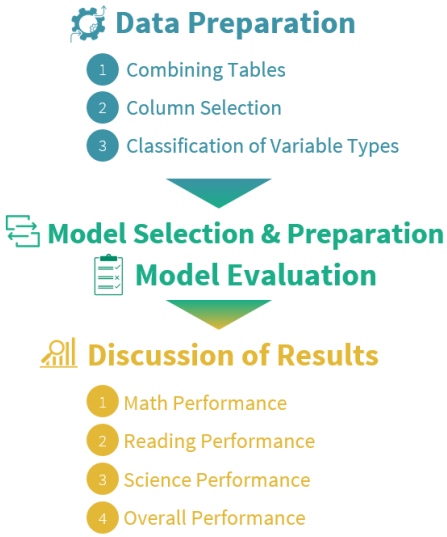


Figure 1. Overview of Methodology

Figure 1 above shows an overview of the analytical processes performed for this paper. Detailed descriptions of each steps are elaborated in the subsequent paragraphs.

DATA PREPARATION

Step 1: Combining Tables

The Standardized Scoring Table, from the scoring standardization performed in Paper XXX-2017, and Student Questionnaire Table from the PISA database, both containing 6,115 records, are combined into a single table using left join.

Step 2: Column Selection

Calculated response columns, such as “PV1SCIE – Plausible Value 1 in Science”, “UNIT - REP_BWGT: RANDOMLY ASSIGNED UNIT NUMBER”, and Warm Likelihood Estimates (WLE) response columns are excluded from our analysis. Only response columns with question terms are kept for our explanatory analysis.

Step 3: Classification of Variable Types

Referencing the Codebook obtained from the OECD PISA 2015 Database, variables are classified into continuous, nominal, and ordinal types. An illustration of the classification of continuous, nominal, and ordinal variables are shown below in Figures 2, 3 and 4 respectively.

ST059Q01TA Number of <class periods> required per week in <test language>	NUM	0 - 40

Figure 2. Example of continuous explanatory variable

ST065Class	Student coded science class (from ST065Q01NA)
1	Physics
2	Chemistry
3	Biology
4	Earth and space
5	Applied sciences and technology
6	General, integrated, or comprehensive science
7	Could not determine

Figure 3. Example of nominal explanatory variable

ST064Q01NA <school science> courses? I can choose the <school science> course(s) I study.	
1	No, not at all
2	Yes, to a certain degree
3	Yes, I can choose freely

Figure 4. Example of ordinal explanatory variable

MODEL SELECTION

When considering an explanatory model for our dataset, we initially considered using both Multiple Linear Regression (MLR) and Partition Models. While performing model fitting using MLR, we notice that the MLR model does not fit well with the dataset used, as there are very few continuous independent variables to draw a meaningful linear relationship with the Standardized Scores of student's. Since most of the independent variables are categorical in nature, we chose to adopt the recursive tree Partition Models for our explanatory analysis.

MODEL PREPARATION

To better understand the differences between top performing students and poor performing ones, we employed the Decision Tree, Bootstrap Forest, and Boosted Tree Partition Methods in JMP Pro 13. A validation factor of 0.3 is used for all three methods. Standardized Scores are assigned to the Response role (Y), and all selected terms from Student Questionnaire are assigned to the Factor role (X). Default options are used for both Bootstrap Forest and Boosted Tree methods.

MODEL EVALUATION

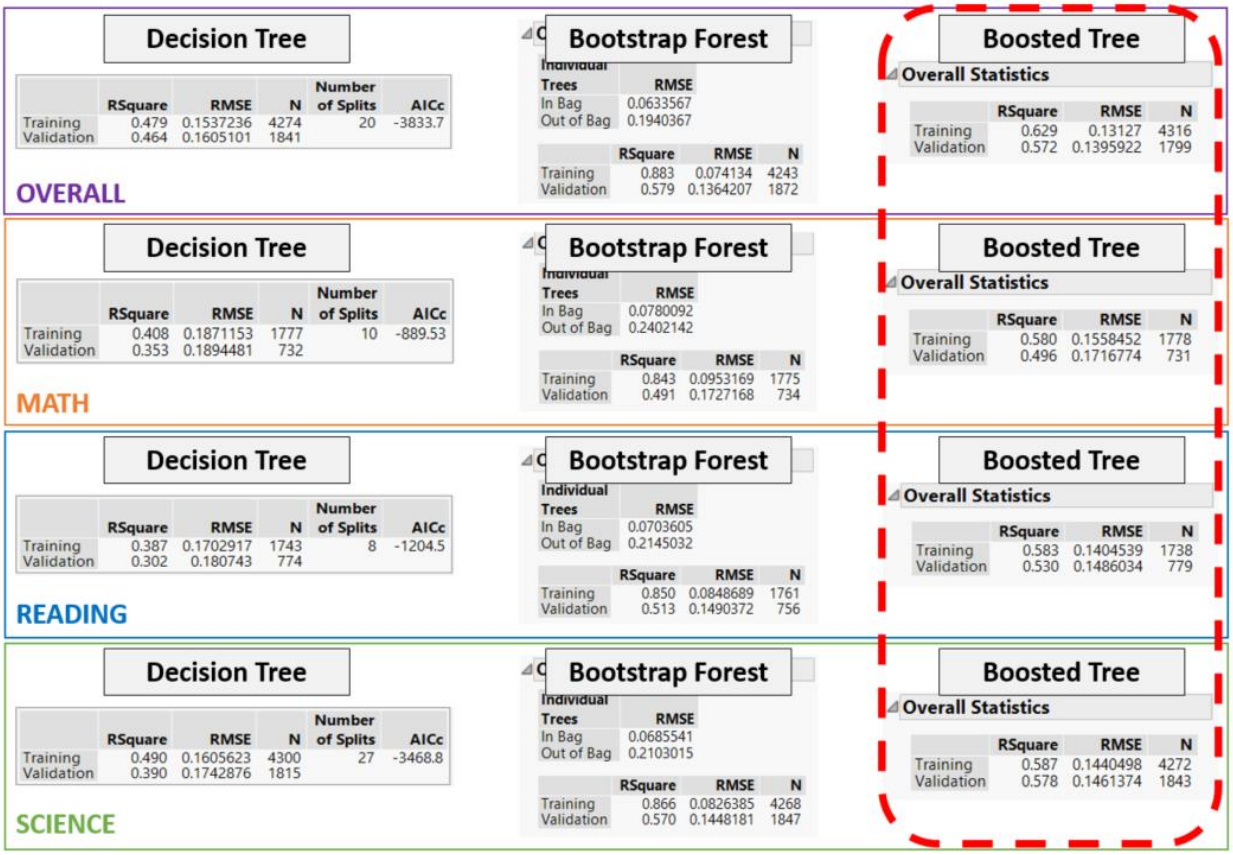
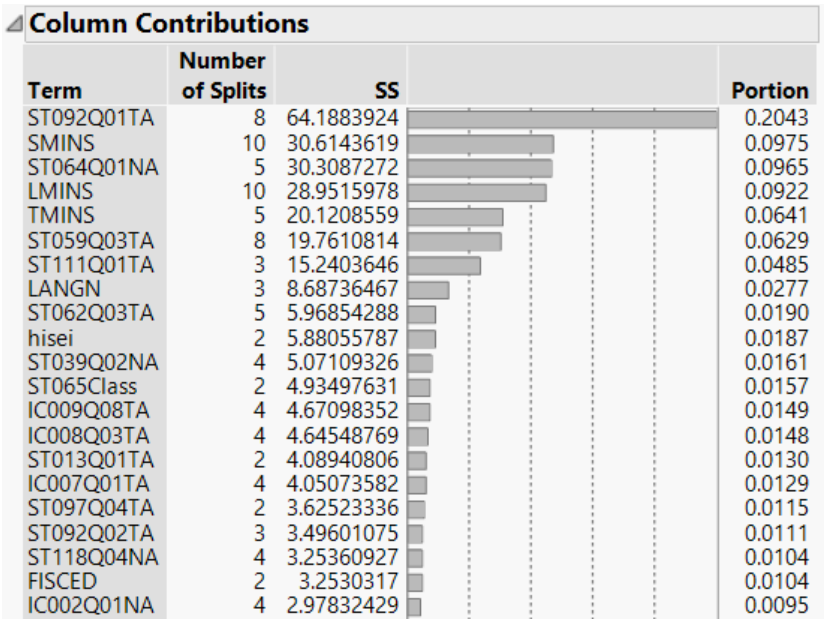


Figure 5. Comparison of Partition Methods

Based on the results shown above in Figure 5, the Boosted Tree is the chosen model for further evaluation, as it generally has a higher Validation RSquared value across the three partitioning models. The difference in RMSE values between Training and Validation sets for the Boosted Tree is also in general, lesser that the other two models, suggesting that the result of the Boosted Tree has less overfitting issues compared to the other two models.

DISCUSSION OF BOOSTED TREE OBSERVATIONS

In order to identify factors with strong influence in student's scores, we only consider Column Contributions with Portion greater than 0.01. Further visualization is performed on selected unique factors for each subjects to discover how the factor influences student's scores. The Column Contribution results for Math, Reading, Science, and Overall scores are reflected in Figures 6, 7, 8 and 9 respectively.

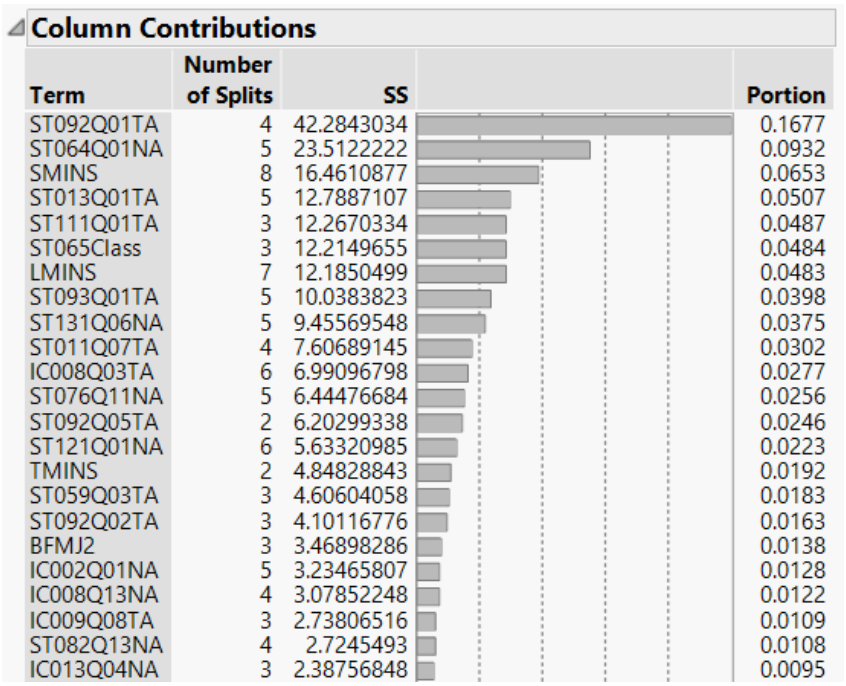


Column Contributions

Term	Number of Splits	SS	Portion
ST092Q01TA	8	64.1883924	0.2043
SMINS	10	30.6143619	0.0975
ST064Q01NA	5	30.3087272	0.0965
LMINS	10	28.9515978	0.0922
TMINS	5	20.1208559	0.0641
ST059Q03TA	8	19.7610814	0.0629
ST111Q01TA	3	15.2403646	0.0485
LANGN	3	8.68736467	0.0277
ST062Q03TA	5	5.96854288	0.0190
hisei	2	5.88055787	0.0187
ST039Q02NA	4	5.07109326	0.0161
ST065Class	2	4.93497631	0.0157
IC009Q08TA	4	4.67098352	0.0149
IC008Q03TA	4	4.64548769	0.0148
ST013Q01TA	2	4.08940806	0.0130
IC007Q01TA	4	4.05073582	0.0129
ST097Q04TA	2	3.62523336	0.0115
ST092Q02TA	3	3.49601075	0.0111
ST118Q04NA	4	3.25360927	0.0104
FISCED	2	3.2530317	0.0104
IC002Q01NA	4	2.97832429	0.0095

Figure 6. Column Contributions for Math Scores

Of the factors having relatively high Column Contributions of above 0.01 for student math scores, we will zoom in on unique factors and their relationship to student's math performance. Factors which will be discussed further in relation to math performance are student's punctuality (ST062Q03TA), student's perception on how they are graded (ST039Q02NA), duration of internet usage over the weekend (IC007Q01TA), how student's feel when studying for a test (ST118Q04NA), as well as their father's education level (FISCED).

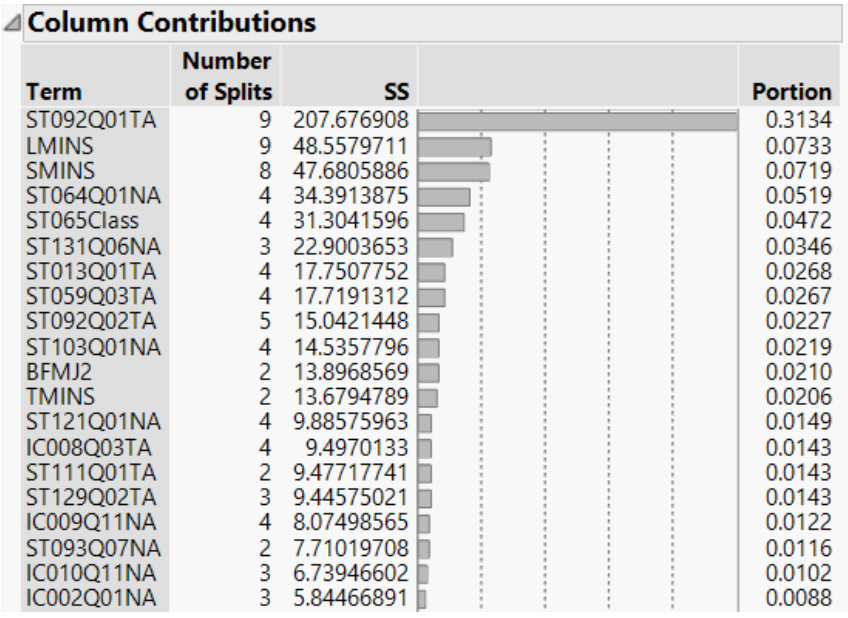


Column Contributions

Term	Number of Splits	SS	Portion
ST092Q01TA	4	42.2843034	0.1677
ST064Q01NA	5	23.5122222	0.0932
SMINS	8	16.4610877	0.0653
ST013Q01TA	5	12.7887107	0.0507
ST111Q01TA	3	12.2670334	0.0487
ST065Class	3	12.2149655	0.0484
LMINS	7	12.1850499	0.0483
ST093Q01TA	5	10.0383823	0.0398
ST131Q06NA	5	9.45569548	0.0375
ST011Q07TA	4	7.60689145	0.0302
IC008Q03TA	6	6.99096798	0.0277
ST076Q11NA	5	6.44476684	0.0256
ST092Q05TA	2	6.20299338	0.0246
ST121Q01NA	6	5.63320985	0.0223
TMINS	2	4.84828843	0.0192
ST059Q03TA	3	4.60604058	0.0183
ST092Q02TA	3	4.10116776	0.0163
BFMJ2	3	3.46898286	0.0138
IC002Q01NA	5	3.23465807	0.0128
IC008Q13NA	4	3.07852248	0.0122
IC009Q08TA	3	2.73806516	0.0109
ST082Q13NA	4	2.7245493	0.0108
IC013Q04NA	3	2.38756848	0.0095

Figure 7. Column Contributions for Reading Scores

Unique factors for student's reading scores, such as their perception on environmental issues (ST093Q01TA), availability of classic literature at home (ST011Q07TA), and their age of exposure to digital devices (IC002Q01NA), will be further analyzed to come up with possible explanations on how these factors potentially influence a student's reading performance score.

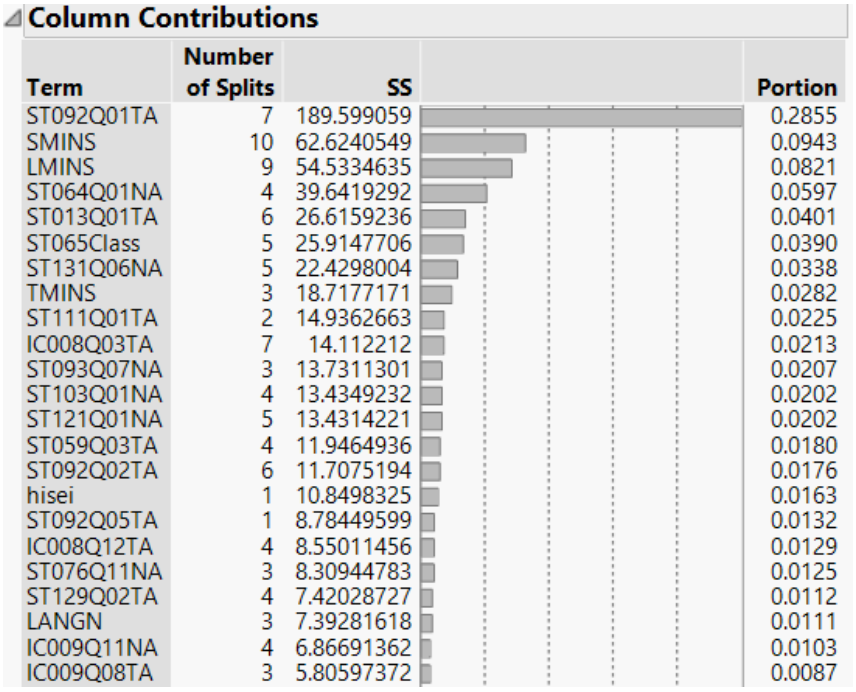


Column Contributions

Term	Number of Splits	SS	Portion
ST092Q01TA	9	207.676908	0.3134
LMINS	9	48.5579711	0.0733
SMINS	8	47.6805886	0.0719
ST064Q01NA	4	34.3913875	0.0519
ST065Class	4	31.3041596	0.0472
ST131Q06NA	3	22.9003653	0.0346
ST013Q01TA	4	17.7507752	0.0268
ST059Q03TA	4	17.7191312	0.0267
ST092Q02TA	5	15.0421448	0.0227
ST103Q01NA	4	14.5357796	0.0219
BFMJ2	2	13.8968569	0.0210
TMINS	2	13.6794789	0.0206
ST121Q01NA	4	9.88575963	0.0149
IC008Q03TA	4	9.4970133	0.0143
ST111Q01TA	2	9.47717741	0.0143
ST129Q02TA	3	9.44575021	0.0143
IC009Q11NA	4	8.07498565	0.0122
ST093Q07NA	2	7.71019708	0.0116
IC010Q11NA	3	6.73946602	0.0102
IC002Q01NA	3	5.84466891	0.0088

Figure 8. Column Contributions for Science Scores

Student's frequency of downloading learning apps on mobile devices (IC010Q11NA) will be studied further to better understand how the downloading of learning apps influences a student's science test scores.



Column Contributions

Term	Number of Splits	SS	Portion
ST092Q01TA	7	189.599059	0.2855
SMINS	10	62.6240549	0.0943
LMINS	9	54.5334635	0.0821
ST064Q01NA	4	39.6419292	0.0597
ST013Q01TA	6	26.6159236	0.0401
ST065Class	5	25.9147706	0.0390
ST131Q06NA	5	22.4298004	0.0338
TMINS	3	18.7177171	0.0282
ST111Q01TA	2	14.9362663	0.0225
IC008Q03TA	7	14.112212	0.0213
ST093Q07NA	3	13.7311301	0.0207
ST103Q01NA	4	13.4349232	0.0202
ST121Q01NA	5	13.4314221	0.0202
ST059Q03TA	4	11.9464936	0.0180
ST092Q02TA	6	11.7075194	0.0176
hisei	1	10.8498325	0.0163
ST092Q05TA	1	8.78449599	0.0132
IC008Q12TA	4	8.55011456	0.0129
ST076Q11NA	3	8.30944783	0.0125
ST129Q02TA	4	7.42028727	0.0112
LANGN	3	7.39281618	0.0111
IC009Q11NA	4	6.86691362	0.0103
IC009Q08TA	3	5.80597372	0.0087

Figure 9. Column Contributions for Overall Scores

Table 1 below shows all the similar factors across all subjects as well as the overall scores. For our analysis, we will focus on questions pertaining to their general knowledge (ST092Q01TA), amount of time spent studying (TMINS), and the level of education the student expects to complete (ST111Q01TA), in relation to their performance in the test.

Compilation of similar factors across subjects

Question No.	Question Description
ST092Q01TA	How informed are you about this environmental issue? The increase of greenhouse gases in the atmosphere
SMINS	Learning time (minutes per week) - <science>
LMINS	Learning time (minutes per week) - <test language>
ST064Q01NA	<school science> courses? I can choose the <school science> course(s) I study.
ST013Q01TA	How many books are there in your home?
ST065Class	Student coded science class (from ST065Q01NA)
TMINS	Learning time (minutes per week) - in total
ST111Q01TA	Which of the following do you expect to complete?
IC008Q03TA	Use digital devices outside school for using email.
ST059Q03TA	Number of <class periods> required per week in <science>
ST092Q02TA	How informed are you about this environmental issue? The use of genetically modified organisms (<GMO>)

Table 1. Compilation of similar factors

MATH PERFORMANCE

Based on the results reflected in the figures below, we noted that students who are more punctual for school tend to perform better (Figure 10), students who feels that they are graded harder compared to other students attained lower scores (Figure 11), students perform better if they use the Internet for between 30 minutes to 4 hours a day (Figure 12), the performance of top and bottom performing students differ in relation to how they feel when studying for a test (Figure 13), and that in general, the higher their Father's education level, the higher their math scores (Figure 14).

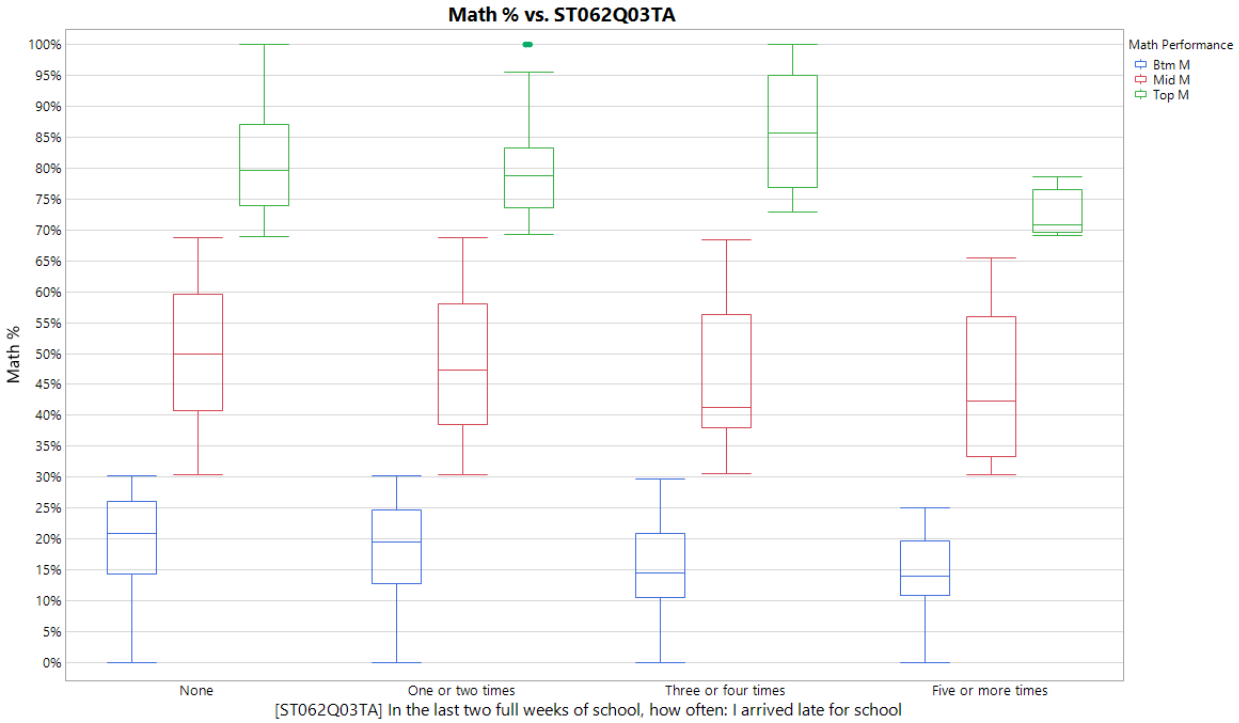


Figure 10. Box Plot of Student's punctuality against Math Scores

From Figure 10, we can deduce that for students with poorer (Mid and Btm) performance in Math, those who arrive late for at most two times score on average 5% more than students who are late for more than three times. An interesting thing to note is that for top performing Math students, those who arrive late for school for three or four times are observed to score on average 5% better than the more punctual students. Also, students who are late for at least five times in two weeks achieved lower scores in comparison with the rest. Understanding their cause of lateness could be a first step in helping them improve their Math performance.

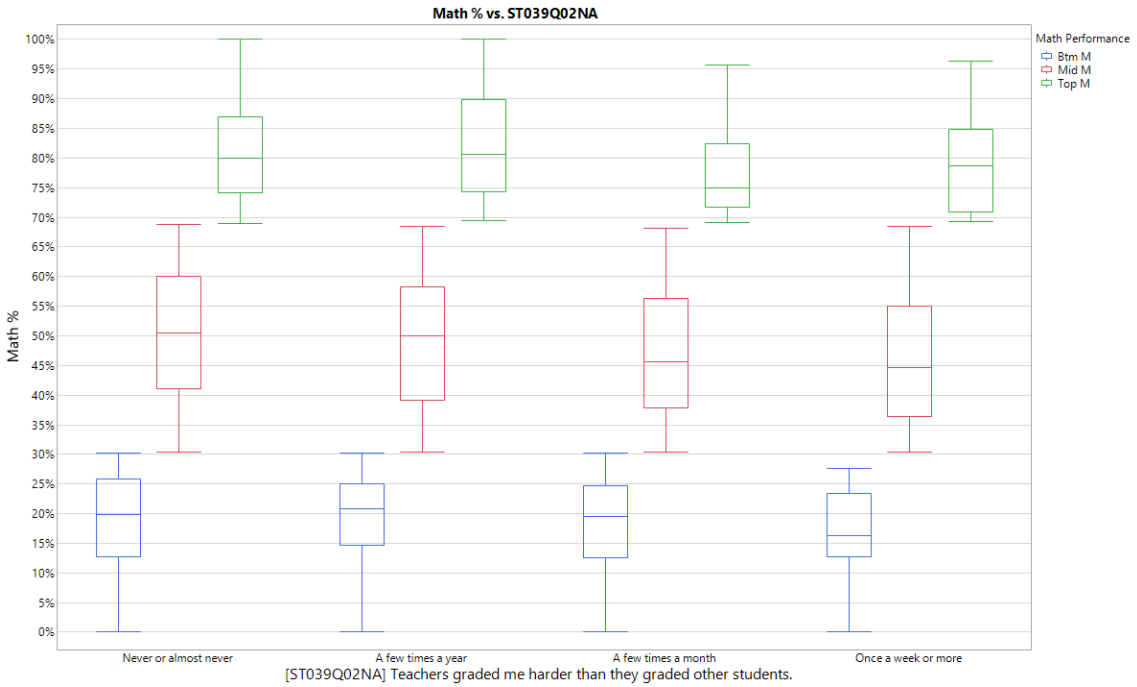


Figure 11. Box Plot of Student's opinion on Teacher's grading against Math Scores

Figure 11 shows that for students with math scores in the bottom 25th percentile, those who report that their teachers seem to grade them harder compared to the other students at least once a week or more generally score around 5% lower than students who view their teacher as fairer in grading their work. To reduce students' perception on how they are graded by their teachers, schools can consider increasing the adoption of technology in grading students' work.

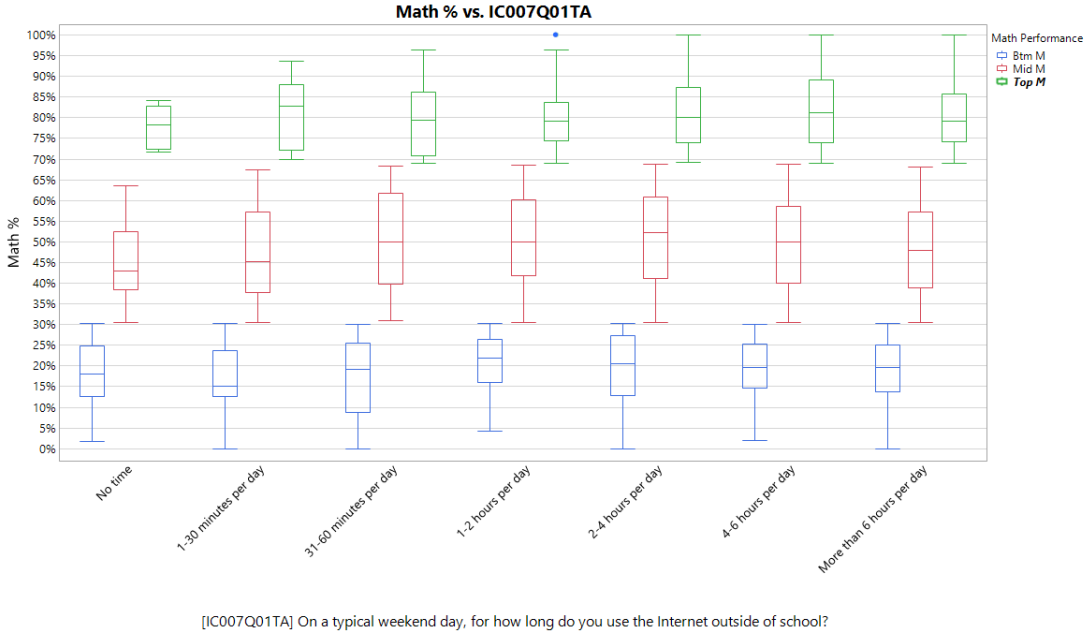


Figure 12. Box Plot of Student's weekend Internet Usage against Math Scores

Surprisingly, as shown in Figure 12, students who rarely use internet over the weekend (less than 30 minutes per day) generally score lower compared to their peers who are on the internet for a longer duration. It is observed that students who use the internet for between 30 minutes to 4 hours per day over the weekend perform better in Math. From this observation, parents could consider allowing more internet usage time instead of prohibiting the use of it, while at the same time monitoring what their child is using the internet for, as a preventive measure.

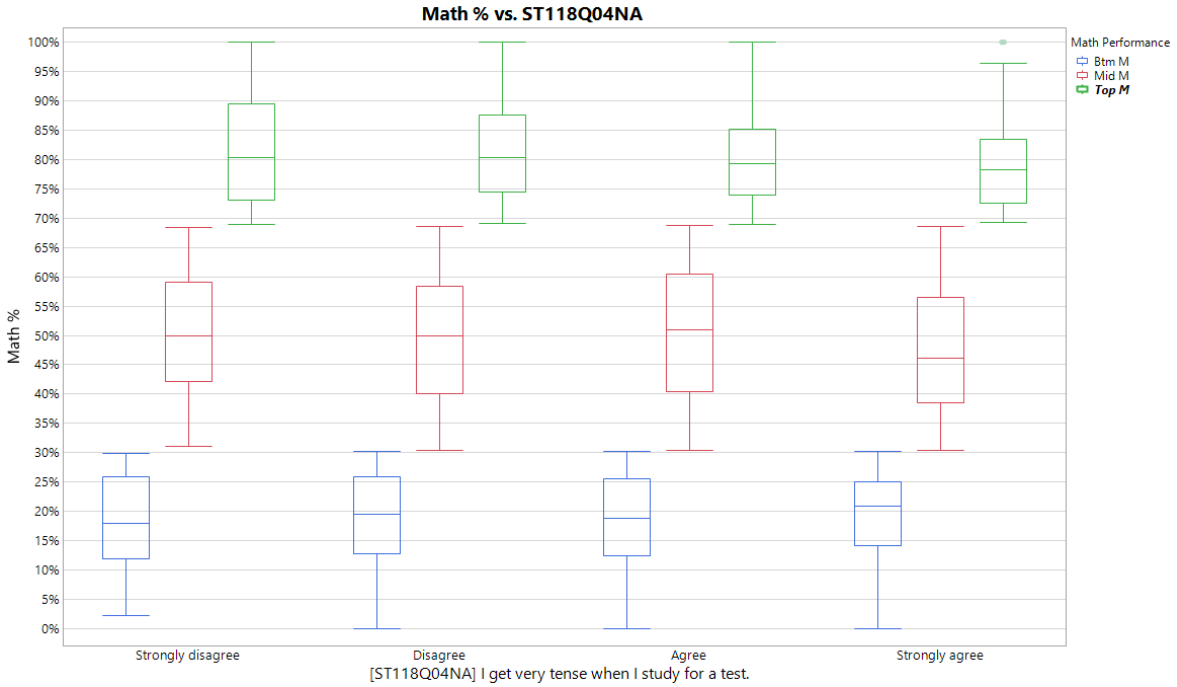


Figure 13. Box Plot of Student's opinion on studying against Math Scores

In relation to Figure 13 above, it is interesting to note that amongst the poorer performing students, the more they feel tense when studying for a test, the better they perform. In contrast, for the top performers, the more tense they feel, the poorer their test scores. A possible explanation for this observation could be that students who are good in the subject feel more confident and prepared when studying for a test, whereas students who are weaker in the subject feel less confident and less prepared, thus feel more tense when they are studying for a test.

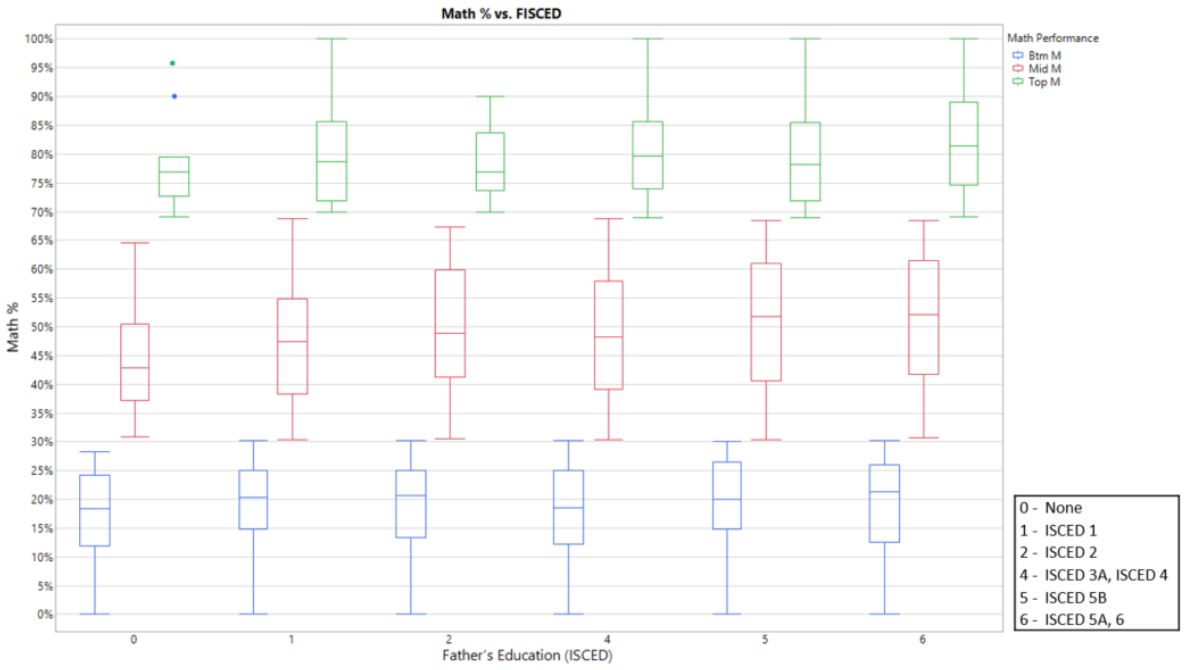


Figure 14. Box Plot of Father's Education level against Math Scores

A general observation from the analysis of a student's father's education level in Figure 14 is that across student performances, the more educated their father is, the better the student performs. This reveals that socioeconomic factors do play an important part in students' test performance.

READING PERFORMANCE

From the visualization of unique factors for reading performance, in the figures shown below, we observed that students who perform better are less optimistic of an improvement in Air Pollution over the next 20 years (Figure 15), has Classic literature available at home (Figure 16), are exposed to digital devices at a young age (Figure 17). These insights reveal that the affluence levels of students as well as exposure to current affairs and environmental issues have a role to play in the performance of a student in the Reading component.

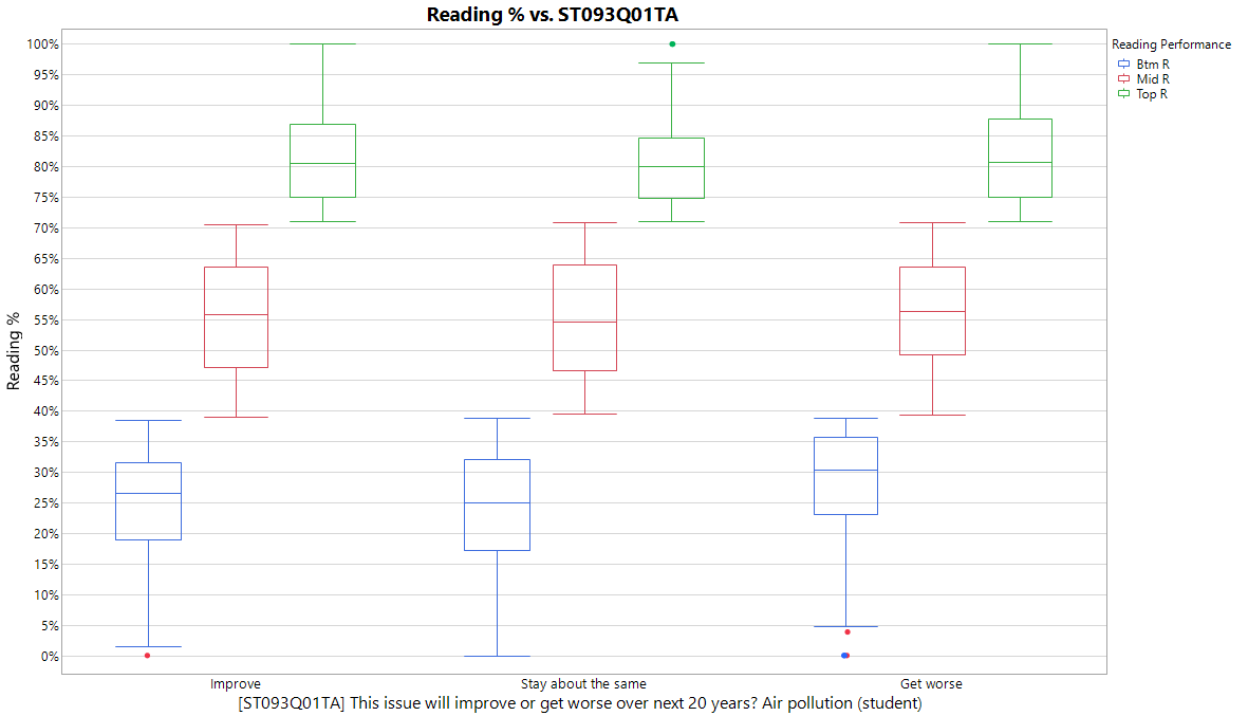


Figure 15. Box Plot of Student's opinion on Air Pollution against Reading Scores

Students who perform better in the Reading component amongst their performance groups, as reflected in Figure 15, are for the opinion that the air pollution issue would get worse over the next 20 years. A possible explanation for this is that this group of students who are less optimistic of an improvement in the air pollution issue, reads widely and are exposed to the negative reports of air pollution around the world.

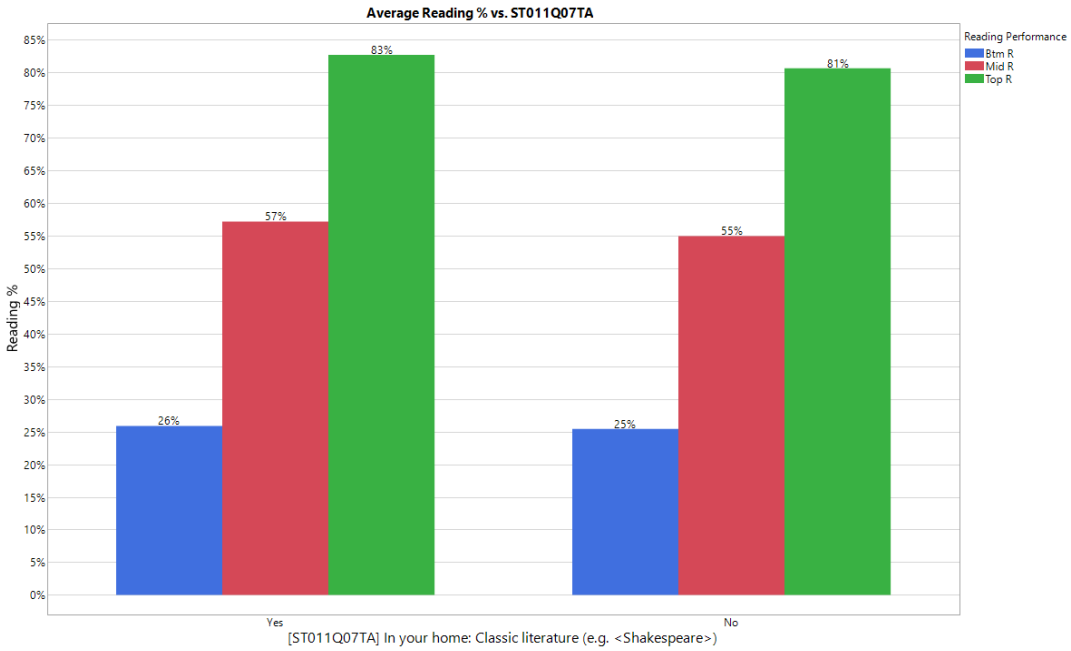


Figure 16. Bar Chart of Classical Literature ownership at home against Reading Scores

Across their performances in the test, students with classic literature at home tend to perform around 2% better than students who do not have classic literature at home, as seen in Figure 16. An implication from this observation is that the type and variety of books available at home could have an influence in student's reading performance in school.

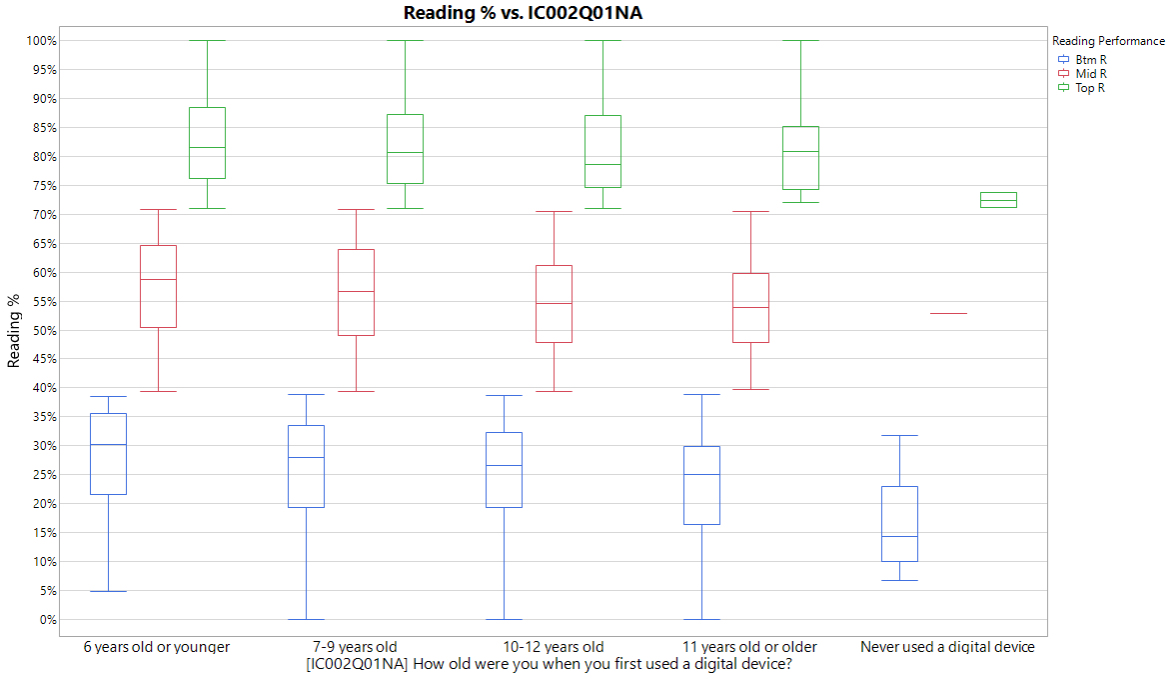


Figure 17. Box Plot of Age of exposure to digital devices against Reading Scores

Figure 17 shows that the younger the students are exposed to digital devices, the better their reading performance. Students who have never used a digital device prior to the test scored significantly lower compared to students in the same performance band. The lack of computer literacy could be the main cause hindering their performance in the test. With more exposure to technology following Singapore's Smart Nation initiative, this issue would be better tackled, and hopefully in the next PISA assessment, the age of exposure to digital device would not be a key factor influencing student's test performance.

SCIENCE PERFORMANCE

In Figure 18 below, students who download learning apps at least once or twice a week are shown to have lower Science scores compared to the less frequent users. This could be due to the differences in syllabus of the learning apps compared to the Science syllabus taught in schools.

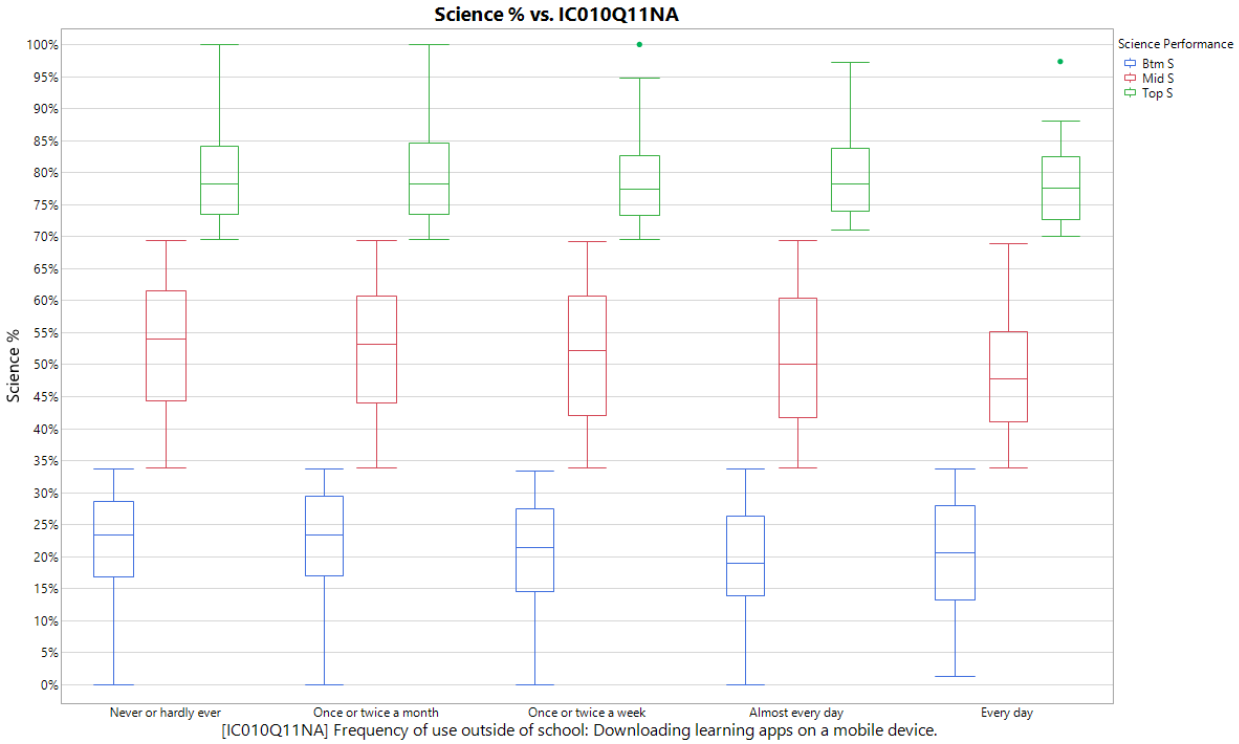


Figure 18. Box Plot of Student's Mobile Device usage outside of school against Science Scores

OVERALL COMPARISONS

Comparing the similar factors across all subjects, student's response to their understanding of greenhouse gases has the most effect on their performance. As shown in Figure 19 below, comparing students in the top 25%, middle 50%, and bottom 25% quantiles, students in each group who have better knowledge of greenhouse gases tend to have higher scores.

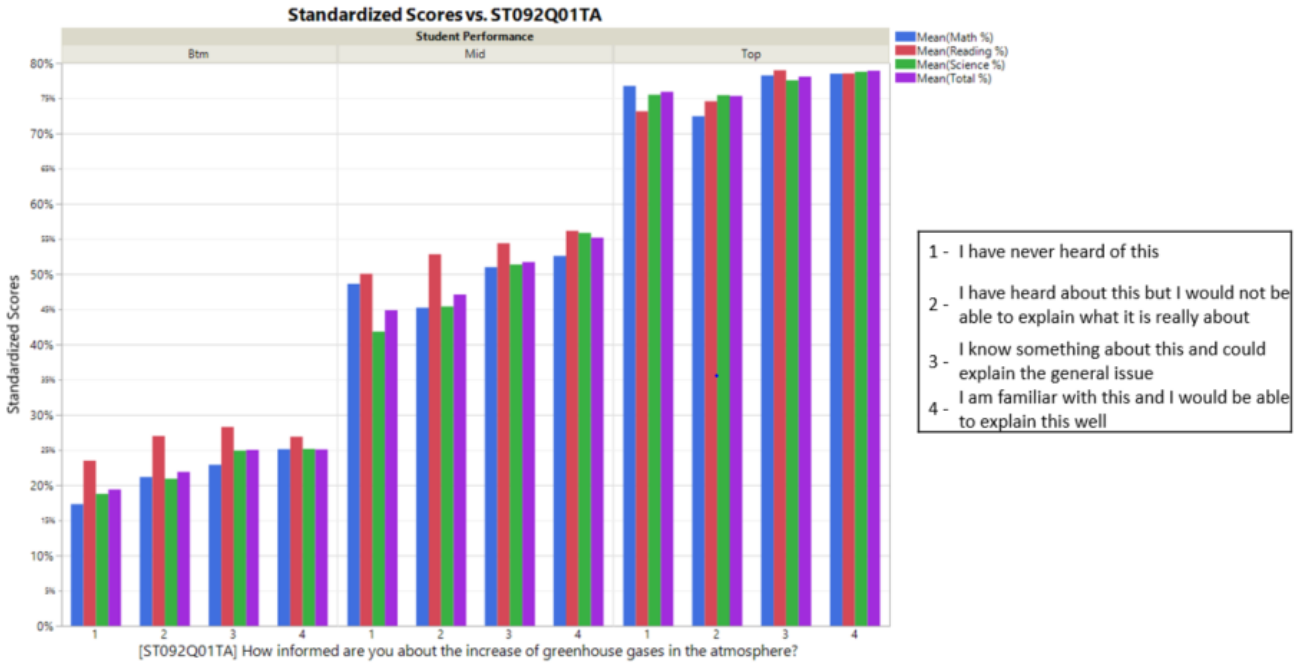


Figure 19. Bar Chart of Student's understanding of greenhouse gases against Student's Scores

Figure 20. Line Plot of Weekly learning time against Student's Scores

When comparing the learning time spent per week among the top, middle, and bottom performing students, the more time spent learning, the better the test scores for middle and bottom performers. In contrast, as the learning time increases, the lower the test scores of top performing students in Reading and Science, and an upward trend for Math scores. From Figure 20 above, the optimal learning time for top performers is observed to be 1800 minutes per week, around 2500 minutes per week for middle performers, and around 2200 minutes per week for the bottom performers.

With regards to the aspirational effects on student's performance, the more aspirational the students are, the better their performance across performance levels, as reflected in Figure 21 below. Students who expect to complete a degree program (6) generally performs on average 3% better than students who expect to complete up to diploma level (5).

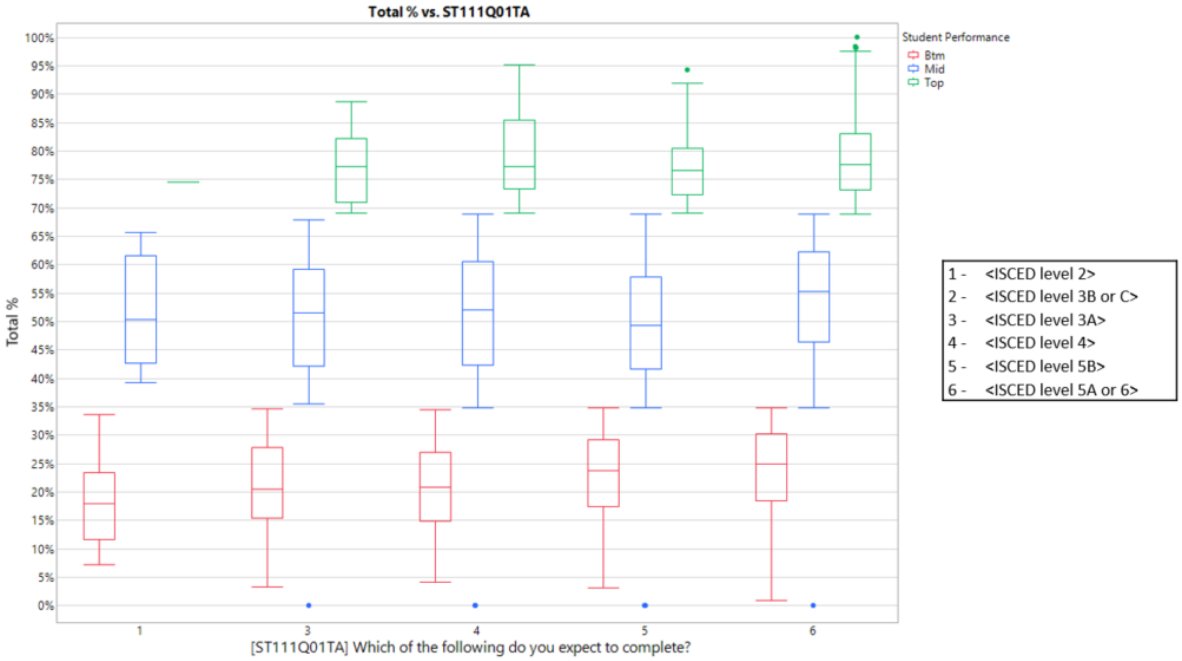


Figure 21. Box Plot of Student's aspirations against Student's Scores

CONCLUSION

In understanding the factors influencing student's performance in Reading, Mathematics, and Science, we learned that socioeconomic factors, such as the Parent's education level, availability of digital devices at home, has a positive effect on student's test scores. In addition, class environment plays a part in the performance of the different groups of students. The perceived fairness of a teacher when grading students' work is seen to have a positively strong influence in student's performance. Student's age of exposure to digital devices is also found to have a strong effect on their test scores, in particular those who have never interacted with digital devices prior to the test performed poorly. In line with the Singapore Government's Smart Nation Masterplan, an early exposure to technology might improve Singapore's performance in the next edition of the PISA Survey.

ACKNOWLEDGMENTS

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] Rai, S., Saini, P., & Jain, A. K. (2014). Decision Tree Algorithm Implementation Using Educational Data. International Journal of Computer-Aided technologies (IJCAx) Vol.1,No.1,April 2014

[2] Sudiya, A. (2016, January). Application Of Decision Tree Approach To Student Selection Model-A Case Study. In IOP Conference Series: Materials Science and Engineering (Vol. 105, No. 1, p. 012014). IOP Publishing.

[3] Al-Barrak, M. A., & Al-Razgan, M. (2016). Predicting students final GPA using decision trees: a case study. International Journal of Information and Education Technology, 6(7), 528.

[4] Quadri, M. M., & Kalyankar, N. V. (2010). Drop out feature of student data for academic performance using decision tree techniques. Global Journal of Computer Science and Technology, 10(2).

[5] Predictive and Specialized Modeling – JMP. Available at https://www.jmp.com/content/dam/jmp/documents/en/support/jmp13/Predictive-and-Specialized-Modeling.pdf

[6] 2015 Technical Report - PISA – OECD. Available at http://www.oecd.org/pisa/data/2015-technical-report/

[7] Changes in the administration and scaling of PISA 2015 and implications for trends analysis – PISA 2015 Results (Volumes I and II). Available at https://www.oecd.org/pisa/data/PISA-2015-vol%202-Annex-A5-Trends-analysis.pdf