Paper ###-2017

Using Latent Class Analysis to Standardize Scores from the Programme for International Student Assessment (PISA) Global Education Survey to Determine Differences between Schools in Singapore

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

ABSTRACT

The Organization 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. A standardized scoring is needed due to the survey using multiple booklets and each booklet contains varying number of questions from a combination of science questions together with reading and/or mathematics questions. We used latent class analysis to determine the difficulty of each question through the profiling of clusters and then adjust the weight of the question based on the difficulty in order to get a standardized score. From the results of LCA and standardization, we have determined that not all schools perform on the same level and there are schools which have performed exceptionally well for the 2015 PISA survey while there are also schools which have performed poorly.

INTRODUCTION

In Singapore, the Ministry of Education (MOE) is responsible for collecting and analyzing data collected from schools in order to improve on their policies and practices in education which they set for schools. However, most of the data from MOE are not publicly available for research and analysis for those who are not working inside MOE. With this limitation, it is hard to gain insights on the performances of schools and its students in Singapore to make improvements or suggestions to the education system. An alternative for deriving insights from Singapore's education system is through the publicly available data collected by the Organization for Economic Co-operation and Development (OECD). The OECD conducts the Programme for International Student Assessment (PISA) global education survey every three years to assess the education systems worldwide through testing 15 year old students in the subjects of mathematics, reading, and science. The 2015 PISA data was released last December 6 2016 and Asian countries continue to perform exceptionally well, with Singapore rated as the best, replacing Shanghai which is now part of a combined entry for China. This paper aims to analyze the recently released PISA data to find out what are the contributing factors resulting in Singapore's stellar performance as well as find out if there are key characteristics or traits for the best performing schools in Singapore.

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." [1] However, being products of the Singapore education system ourselves, we acknowledge that this not seem to be that case, and there are still disparities that exist. We aim to analyze the performances of students based on their socio-economic background and see if this plays a part. Furthermore, Singapore's Minister for Education, Mr. Heng Swee Keat also started an "every school a good school" slogan [2]. Therefore, the focus of this paper is to also find out if it is fair to state that all schools are the same based on the PISA data that we will be using.

The rest of the paper is organized as follows. Following the introduction, we review the literatures related to our study which is on latent class analysis. This is followed by an overview on our methodology and our data preparation steps. The next part of the paper will focus on the method used followed by an analysis of the results. With the results, the standardized scores can be obtained and insights can be derived. Lastly, the paper concludes by highlighting the key findings from the method and analysis.

LITERATURE REVIEW

An analytics practicum team did an analysis on the 2012 PISA data in 2016[3]. The team did an aggregation of student scores by dividing the total number of correct answers by the total number of questions the student attempted and then calculated the average test score for each school and every subject based on the percentage score of students. The team had an assumption that school, family and personal factors can all affect the performance of a student and used techniques such as K-means clustering, partition analysis, and regression model for their analysis and variable selection. The team was able to identify factors which affect the performance of students and made recommendations on how schools can identify high risk students based on various factors.

Latent class analysis (LCA) is a statistical method for finding subtypes of related cases (latent classes) from multivariate categorical data. The results of LCA can be used to classify cases to their most likely latent classes. Common areas for the use of LCA are in health research, marketing research, sociology, psychology, and education [4]. This clustering algorithm offers several advantages over traditional clustering approaches such as K-means such as assigning a probability to the cluster membership for each data point instead of relying on the distances to biased cluster means and LCA provides various diagnostic information such as common statistics, Bayesian information criterion (BIC), and p-value to determine the number of clusters and the significance of the variables' effects [5].

This method was applied on the 2012 PISA data of Taiwan to objectively classify students' learning strategies to determine the optimal fitting latent class model of students' performance on a learning strategy assessment and to explore the mathematical literacy of students who used various learning strategies. The findings of the research shows that a four class model was the optimal fit model of learning strategy based on the BIC and adjusted BIC when comparing the four class model to other models of two to five classes. The study showed that Taiwanese students who were classified under the "multiple strategies" and "elaboration and control strategies" group (multiple learning strategy) tend to score higher than average while students classified under the "memorization" and "control" group (single learning strategy) performed lower than average [6].

METHODOLOGY

Figure 1. Flow diagram illustrating analytical process

Figure 1 illustrates the analytical process used for this paper. Multiple steps are done for the data preparation as the data is directly obtained from the PISA database where the 2015 PISA data was released last December 6 2016. We then proceed with the data analysis using LCA to determine the optimal fitting latent class model and a discussion on the results from the method. From the results of LCA, every question's weight will be adjusted to determine the standardized score of each student.

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

From the PISA 2015 Database, we only used the files relevant for the project which are the student questionnaire data, school questionnaire data, and cognitive item data. The other files which were not relevant for us were the teacher questionnaire data and the questionnaire timing data. We also used the codebook data file for easy reference.

Upon initial exploration of the cognitive item data, which contains information on how students answered mathematics, reading, and science questions, we noticed that there were multiple booklets used and discovered a pattern. Booklets 31 to 96 were used for schools in Singapore and all booklets contained questions for Science together with Reading and/or Mathematics questions or just purely Science questions. Each booklet contained various number of questions and thus the total scores of each student cannot be compared across booklets. LCA will be used to determine the difficulty of the each question based on how well the students performed for the question and then the questions will be adjusted based on the difficulty.

Booklet ID	Reading	Math	Science
31 - 42	✓		✓
43 - 54		✓	✓
55 - 66	✓	✓	✓
67 - 78		✓	✓
79 - 90	✓		✓
91 - 96			✓

Figure 2. Breakdown of booklets based on distribution of reading, math, and science questions

DATA PREPARATION

For steps 1 to 4, we used Excel in preparing the data while JMP Pro was used for the rest of the steps.

Step 1:

From the raw files extracted from the PISA 2015 database, we only kept those with the 3 character country code of "SGP" as we only want the data related to Singapore. This is applied to the student questionnaire data, teacher questionnaire data and cognitive item data. This provided us with 6115 students and 177 schools of which 168 are public schools while the other 9 are private schools.

Step 2:

The next step was to remove columns with no responses from all schools and students. Columns that contained the same value in all entries were also removed such as Region and OECD Country. For the student questionnaire data and teacher questionnaire data, this is the last step for data preparation while more steps are needed for the cognitive item data in terms of having a standardized score for each student.

Step 3:

In the cognitive item data, each question contained several information such as raw response, scored response, timing, and number of actions. The only columns which were kept for the cognitive item data were the scored responses or coded responses as this contains the information on whether the student received any points for the question.

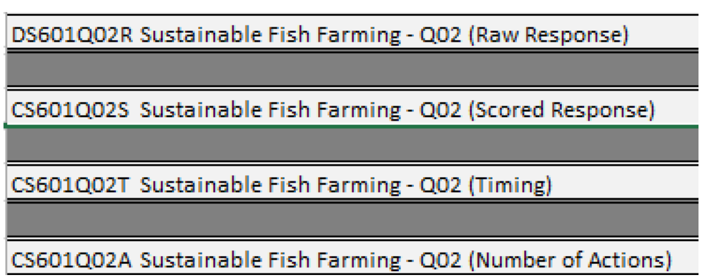


Figure 3. Information for each questions (Step 3)

Step 4:

Questions in the cognitive item data were scored differently as some questions were given the value of 1 for partial credit and 2 for full credit. We decided to allocate 0 for no credit, 0.5 for partial credit, and 1 for full credit. For missing values, the value of 9999 was given.

0	0 - No credit
1	1 - Partial credit
2	2 - Full credit
6 / .R	Not Reached
7 / .N	Not Applicable
8 / .I	Invalid
9 / .M	No Response
SYSTEM MISSING	Missing

Figure 4. Value of 1 for partial credit and 2 for full credit

0	00 - No credit
11	11 - Partial credit
12	12 - Partial credit
13	13 - Partial credit
21	21 - Full credit
22	22 - Full credit
23	23 - Full credit
96 / .R	Not Reached
97 / .N	Not Applicable
99 / .M	No Response
SYSTEM MISSING	Missing

Figure 5. Multiple values for partial credit and full credit

Step 5:

In Excel, each student belonged to a row and the columns contained the student's score for the questions that were answered. We then transposed the questions from columns to rows in JMP Pro to get the count and distribution of scoring classification for each question.

	CNTSCHID	STRATUM	CNTSTUID	CBASCI	BOOKID	Question	Subject	N(0)	N(0.5)	N(1)	N(9999)
1	70200001	SGP0203	70200112	4	31	DR219Q01AC	R	0	0	1	0
2	70200001	SGP0203	70200112	4	31	DR219Q01BC	R	0	0	1	0
3	70200001	SGP0203	70200112	4	31	DR219Q01CC	R	0	0	1	0
4	70200001	SGP0203	70200112	4	31	DR219Q01DC	R	0	0	1	0
5	70200001	SGP0203	70200112	4	31	DR219Q01EC	R	0	0	1	0
6	70200001	SGP0203	70200112	4	31	DR219Q01C	R	0	0	1	0
7	70200001	SGP0203	70200112	4	31	DR219Q02C	R	0	0	1	0
8	70200001	SGP0203	70200112	4	31	CR067Q01S	R	0	0	1	0
9	70200001	SGP0203	70200112	4	31	DR067Q04C	R	0	0	1	0

Figure 6. Transposed question from columns to rows in JMP Pro

As there were different booklets used, each student only answered a small portion of all the questions that were available. When the cognitive item data was transposed into JMP Pro, the questions which were not in the booklet answered by the students were also included and this gave a total of 2,109,675 rows. After removing rows which did not contain any value in N(0), N(0.5), N(1), and N(9999), we were left with 314,366 rows.

Step 6:

We then proceeded to bin the scoring classifications based on its distribution for every question. 10 bins were used of 10% ranges. The data is then ready for LCA after this step.

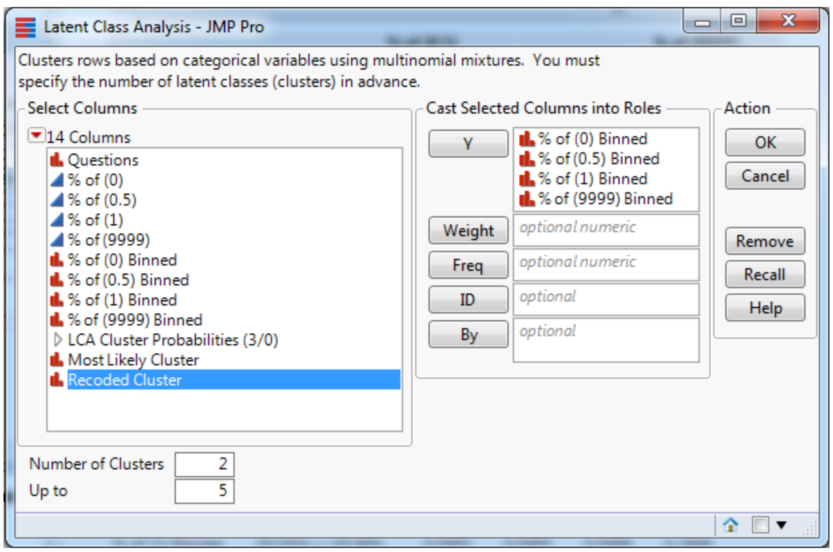
	Questions	% of (0)	% of (0.5)	% of (1)	% of (9999)	% of (0) Binned	% of (0.5) Binned	% of (1) Binned	% of (9999) Binned
1	DR219Q01AC	5.12%	0.00%	87.65%	7.23%	0.00%—10.00%	0.00%—10.00%	80.00%—90.00%	0.00%—10.00%
2	DR219Q01BC	2.23%	0.00%	90.67%	7.10%	0.00%—10.00%	0.00%—10.00%	90.00%—100.00%	0.00%—10.00%
3	DR219Q01CC	1.97%	0.00%	91.72%	6.31%	0.00%—10.00%	0.00%—10.00%	90.00%—100.00%	0.00%—10.00%
4	DR219Q01DC	6.70%	0.00%	86.47%	6.83%	0.00%—10.00%	0.00%—10.00%	80.00%—90.00%	0.00%—10.00%
5	DR219Q01EC	14.19%	0.00%	78.19%	7.62%	10.00%—20.00%	0.00%—10.00%	70.00%—80.00%	0.00%—10.00%
6	DR219Q01C	14.59%	0.00%	79.63%	5.78%	10.00%—20.00%	0.00%—10.00%	70.00%—80.00%	0.00%—10.00%
7	DR219Q02C	14.06%	0.00%	81.60%	4.34%	10.00%—20.00%	0.00%—10.00%	80.00%—90.00%	0.00%—10.00%
8	CR067Q01S	10.51%	0.00%	89.09%	0.39%	10.00%—20.00%	0.00%—10.00%	80.00%—90.00%	0.00%—10.00%
9	DR067Q04C	19.45%	33.51%	44.15%	2.89%	10.00%—20.00%	30.00%—40.00%	40.00%—50.00%	0.00%—10.00%
10	DR067Q05C	20.76%	8.94%	66.75%	3.55%	20.00%—30.00%	0.00%—10.00%	60.00%—70.00%	0.00%—10.00%
11	DR102Q04C	70.70%	0.00%	22.21%	7.10%	70.00%—80.00%	0.00%—10.00%	20.00%—30.00%	0.00%—10.00%
12	DR102Q05C	43.89%	0.00%	53.48%	2.63%	40.00%—50.00%	0.00%—10.00%	50.00%—60.00%	0.00%—10.00%

Figure 7. Binning of scoring classifications

DATA ANALYSIS

Latent Class Analysis:

With the binned scoring classifications, latent class analysis can be performed to determine the most likely difficulty of the questions. To determine the number of clusters to be used, a selection of 2 to 5 clusters was chosen to determine the best fit to the data. The Bayesian information criteria (BIC) was looked at in order to determine the best model fit.



Display 1. Latent class analysis window with selected columns and range of clusters

From the results, the Bayesian information criteria (BIC) was looked at for two to five clusters and the lowest value was determined to be the model with the best fit. From table 1, we could see that the latent class analysis with 4 clusters provided the best fit with a BIC value of 2921.4. However, the latent class analysis with 3 clusters also provided a low BIC value of 2927.15 and thus, we decided to use the latent class analysis with 3 clusters to signify the 3 difficulties for the questions which are easy, medium, and hard.

Latent Class Model	BIC statistic
2 Clusters	2990.5
3 Clusters	**2927.15**
4 Clusters	**2921.4**
5 Clusters	2954.2

Table 1. Latent class models with BIC statistic

Discussion on Latent Class Analysis Results:

From the transposed parameter estimates, the probability of each question's most likely difficulty can be determined based on the conditional probabilities of each cluster. From cluster 1, we can see that the biggest contributor comes from 60.00% to 80.00% of % of (1) Binned which are questions where students have gotten full marks and the second biggest contributor comes from 20.00% to 40.00% of % of (0) Binned which are questions where students have gotten no marks. From the contributors we can profile this cluster to be questions which have medium difficulty. For cluster 2, the biggest contributor comes from 80.00% to 100.00% of % of (1) Binned and 0.00% to 20.00% of % of (0) Binned which means these are questions where students generally get full marks and thus we can profile this cluster to be questions which have easy difficulty. For cluster 3, the biggest contributors are 0.00% to 50.00% of % of (1) Binned, 40.00% to 90.00% of % of (0) Binned, and 10.00% to 40.00% of % of (0.5) Binned. We can profile this cluster to be questions which have hard difficulty since these are questions where more students get no marks or only partial marks.

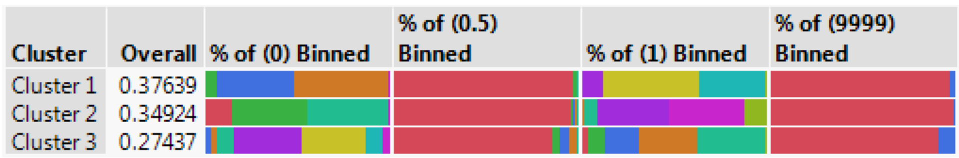


Figure 8. Latent Class Analysis of 3 cluster model

Transposed Parameter Estimates

Column	Category	Cluster 1	Cluster 2	Cluster 3
% of (0) Binned	0.00%—10.00%	0.0003	0.1493	0.0004
% of (0) Binned	10.00%—20.00%	0.0687	0.4064	0.0008
% of (0) Binned	20.00%—30.00%	0.4151	0.0003	0.0320
% of (0) Binned	30.00%—40.00%	0.5145	0.0008	0.0320
% of (0) Binned	40.00%—50.00%	0.0004	0.4419	0.0913
% of (0) Binned	50.00%—60.00%	0.0003	0.0004	0.3687
% of (0) Binned	60.00%—70.00%	0.0003	0.0003	0.3478
% of (0) Binned	70.00%—80.00%	0.0003	0.0003	0.0951
% of (0) Binned	80.00%—90.00%	0.0003	0.0003	0.0320
% of (0.5) Binned	0.00%—10.00%	0.9740	0.9637	0.8628
% of (0.5) Binned	10.00%—20.00%	0.0244	0.0101	0.0401
% of (0.5) Binned	20.00%—30.00%	0.0005	0.0088	0.0533
% of (0.5) Binned	30.00%—40.00%	0.0005	0.0086	0.0431
% of (0.5) Binned	40.00%—50.00%	0.0005	0.0088	0.0007
% of (1) Binned	0.00%—10.00%	0.0003	0.0003	0.0319
% of (1) Binned	10.00%—20.00%	0.0003	0.0003	0.0950
% of (1) Binned	20.00%—30.00%	0.0003	0.0003	0.1793
% of (1) Binned	30.00%—40.00%	0.0003	0.0101	0.3247
% of (1) Binned	40.00%—50.00%	0.0003	0.0761	0.3671
% of (1) Binned	50.00%—60.00%	0.1160	0.3886	0.0005
% of (1) Binned	60.00%—70.00%	0.5225	0.0003	0.0004
% of (1) Binned	70.00%—80.00%	0.3596	0.0021	0.0004
% of (1) Binned	80.00%—90.00%	0.0003	0.4058	0.0004
% of (1) Binned	90.00%—100.00%	0.0003	0.1161	0.0004
% of (9999) Binned	0.00%—10.00%	0.9757	0.9986	0.9141
% of (9999) Binned	10.00%—20.00%	0.0243	0.0014	0.0859

Figure 9. Transposed Parameter Estimates

Cluster	Difficulty
Cluster 1	Medium
Cluster 2	Easy
Cluster 3	Hard

Table 2. Cluster with difficulty level based on profiling

Standardized Scoring

From the latent class analysis, 3 columns are created in the data table with the binned scoring classifications which are the probability of each difficulty (easy, medium, and hard). From these 3 columns, the most likely cluster is derived based on the column with the highest probability.

Prob in Medium	Prob in Easy	Prob in Hard	Most Likely Cluster
8.2362003e-7	0.9999973249	1.8515159e-6	Easy
2.7115109e-6	0.9999889343	8.3541687e-6	Easy
2.7115109e-6	0.9999889343	8.3541687e-6	Easy
8.2362003e-7	0.9999973249	1.8515159e-6	Easy
0.0000500252	0.9999309148	0.0000190599	Easy
0.0000500252	0.9999309148	0.0000190599	Easy
0.0000394369	0.999941914	0.0000186491	Easy
0.0000394369	0.999941914	0.0000186491	Easy
0.9866573206	0.0013065138	0.0120361656	Medium
0.9942583343	0.0057390453	2.6203851e-6	Medium
4.864051e-6	5.9006039e-6	0.9999892353	Hard
0.9999262397	5.7220175e-7	0.0000731881	Medium
0.0000394369	0.999941914	0.0000186491	Easy
1.3304652e-6	1.6132772e-6	0.9999970563	Hard
0.9999428389	1.3412866e-6	0.0000558198	Medium
0.9999262397	5.7220175e-7	0.0000731881	Medium

Figure 10. Probability of difficulty based from the latent class analysis

Each question's weight is then adjusted based on the difficulty of the question and the total score for each student can then be computed for. With the adjusted total score for every student, each school's performance can be calculated for based on the scores of the students.

School ID	STRATUM	School Mean (Math %)	School Mean (Reading %)	School Mean (Science %)	School Mean (Total %)
70200001	SGP0203	60.327%	72.664%	66.111%	66.209%
70200002	SGP0101	82.029%	77.269%	81.922%	81.663%
70200003	SGP0203	47.171%	61.057%	50.525%	51.795%
70200004	SGP0101	25.995%	47.830%	36.972%	37.990%
70200005	SGP0101	35.013%	42.179%	39.520%	39.390%
70200006	SGP0201	51.653%	51.703%	52.888%	53.961%
70200007	SGP0201	54.584%	50.998%	50.467%	50.698%
70200008	SGP0101	33.607%	48.771%	37.491%	38.570%
70200009	SGP0201	41.706%	46.598%	49.531%	49.055%
70200010	SGP0101	42.838%	51.260%	46.709%	45.118%
70200011	SGP0101	36.066%	40.715%	36.538%	36.664%
70200012	SGP0201	52.913%	51.825%	54.129%	53.975%
70200013	SGP0201	44.184%	44.317%	45.293%	44.489%
70200014	SGP0201	38.958%	42.396%	45.715%	45.291%
70200015	SGP0201	50.443%	46.251%	55.341%	52.941%
70200016	SGP0101	21.912%	46.032%	23.325%	24.692%
70200018	SGP0101	22.582%	49.456%	41.369%	40.228%

Figure 11. Mean scores for every subject and overall mean score

DATA EXPLORATION

From the distribution report, we took a look at the mean score distribution for all 177 schools and it can be seen that not all schools performed the same. The mean score was 51.28% for all schools and the middle 50% of schools had a range score of 41.23% to 58.39%.

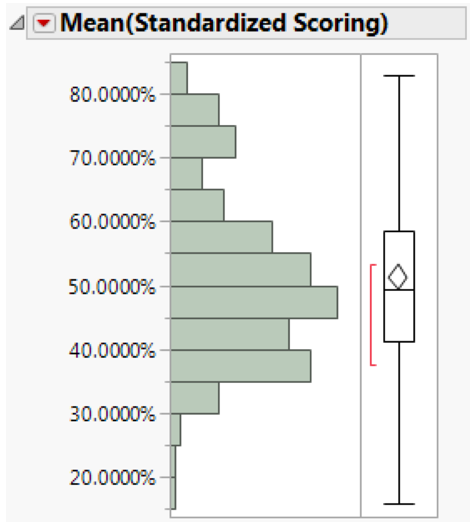


Figure 12. Distribution of mean score

Quantiles

100.0%	maximum	82.8971%
99.5%		82.8971%
97.5%		78.8448%
90.0%		72.2012%
75.0%	quartile	58.3913%
50.0%	median	49.2170%
25.0%	quartile	41.2267%
10.0%		36.5593%
2.5%		29.7756%
0.5%		15.7393%
0.0%	minimum	15.7393%

Summary Statistics

Mean	0.5128268
Std Dev	0.1313274
Std Err Mean	0.0098712
Upper 95% Mean	0.5323079
Lower 95% Mean	0.4933457
N	177

Figure 13. Quantiles and summary statistic

Looking at the boxplot of the scores of all schools, we see how schools in Singapore are different in terms of their performance. There are schools which perform exceptionally well as seen in the right side of the image below while there are also schools which did not perform well which is contrary to the notion of every school being a good school. Another point to highlight in the box plot is the number of outliers for schools which performed well. Although majority of the students in the high performing schools did well, there are a lot more outliers compared to schools in the middle and bottom tier.

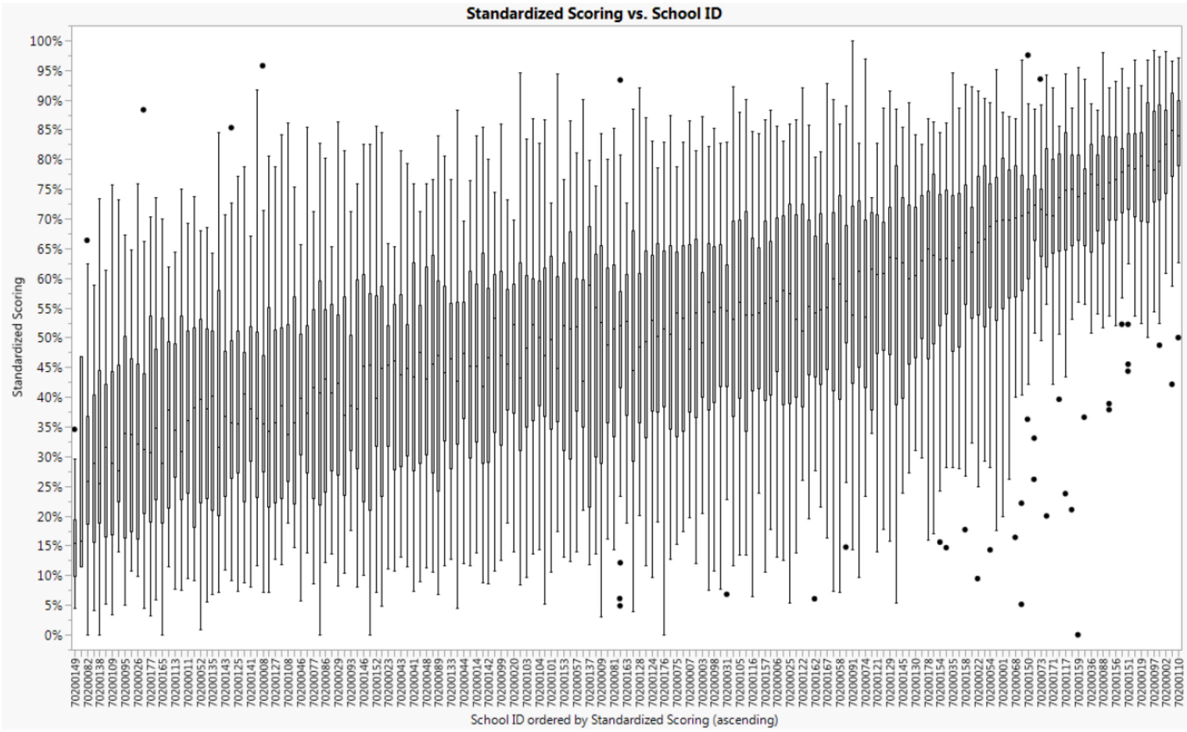


Figure 14. Boxplot of each school's performance

We also took a look at whether there were differences between the students of public and private schools. We performed an analysis of variance on the data with the response being the standardized score of the students and the grouping by the school category which is either public or private. From the results, there is a significant difference between the performance of public school students and private school students as seen in figure 13. To further support this, the p-value in figure 14 is 0.0009 which supports figure 13 that there are significant differences between the students from public schools and private schools.

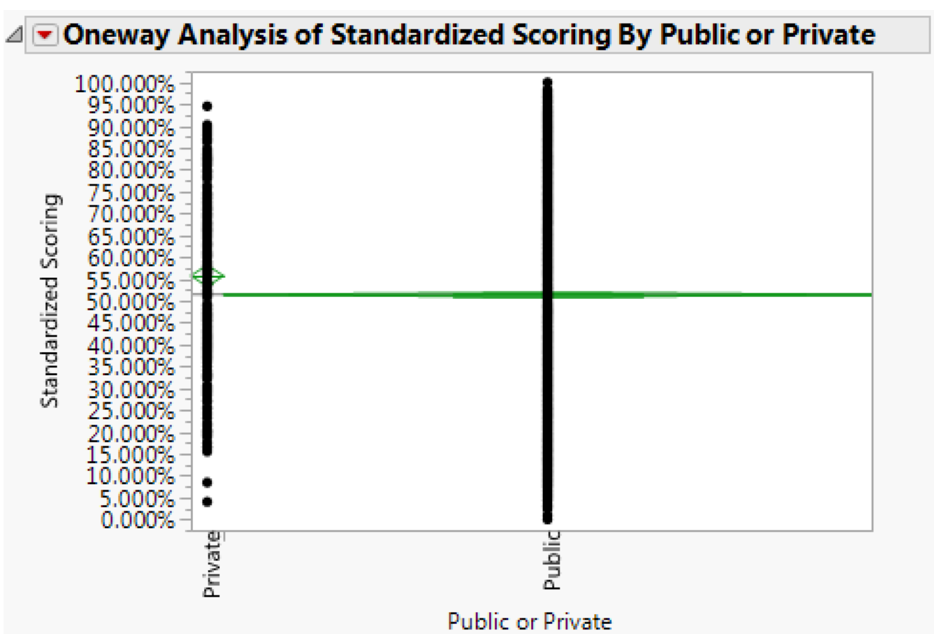


Figure 15. One way analysis

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Public or Private	1	0.51273	0.512735	11.1130	0.0009*
Error	6113	282.04463	0.046138		
C. Total	6114	282.55736			

Figure 16. Analysis of variance

CONCLUSION

With the use of latent class analysis, a three-class model and four-class model was found to provide the best fit to our data as both had a low BIC value of 2927.15 and 2921.4 respectively. The three-class model was picked over the four-class model to signify the three levels of difficulty which were easy, medium, and hard. The three clusters from the model were then profiled based on the parameter estimates to determine the associated difficulty of the cluster. We were then able to standardize the scores for each student based on the results of the latent class analysis due to the fact that there were numerous booklets used in all the schools and each booklet contained different number of questions with varying levels of difficulty. We also discovered that not all students had to answer questions from all 3 subjects which were mathematics, reading, and science due to the use of various booklets and thus, standardization of scores were needed in order to compare performances of students and schools with one another. From the data exploration, it can be seen that there are indeed differences between schools in Singapore when viewing the performances of all schools through a boxplot. We also found out that the private schools performed better than public schools through an analysis of variance. Thus, the slogan of “every school a good school” used by Singapore’s Minister for Education, Mr. Heng Swee Keat, is not true as there are schools which performed poorly compared to other schools. The Programme for International Student Assessment (PISA) global education survey allowed us to discover that there are differences between performances of schools but this data is only conducted on a small percentage of students in Singapore and in order to improve on the performances of schools, a wider sample is needed to have a better analysis on the performance of schools in Singapore.

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] Coughlan, Sean. “Pisa tests: Singapore top in global education rankings”. BBC. 6 December 2016. Available at http://www.bbc.com/news/education-38212070.

[2] Davie, Sandra. “Heng Swee Keat as education minister: A study in bold moves”. Straits Times. 12 October 2015. Available at http://www.straitstimes.com/singapore/education/heng-swee-keat-as-education-minister-a-study-in-bold-moves.

[3] Poh, Jin Hui, Nguyen, Le Hong Ngoc, Zhao, Yazhi. “An Analysis of Singapore student performance in the OECD Programme for International Student Assessment”. Available at https://wiki.smu.edu.sg/ANLY482/img_auth.php/6/6f/Group_15_final_report.pdf.

[4] Uebersax, John. “LCA Frequently Asked Questions (FAQ)”. 8 July 2009. Available at http://www.john-uebersax.com/stat/faq.htm#whatis.

[5] Xu, Beijie. “Clustering Educational Digital Library Usage Data: Comparison of Latent Class Analysis and K-Means Algorithm”(2011). *All Graduate Theses and Dissertations.* Paper 954.

[6] Lin, Su-Wei, Tai, Wen-Chun. “Latent Class Analysis of Students’ Mathematics Learning Strategies and the Relationship between Learning Strategy and Mathematical Literacy”(2015). Universal Journal of Educational Research 3(6): 390-395.