



SHOW ME THE NUMBERS

ISSS602 Data Analytics Lab_Assign1



SEPTEMBER 3, 2016

ZHAO JIANZHANG

MITB-AT

1. Executive Report

1.1 Brief Summary of Data Analysis

The overall performance of 15-year-old Singaporean students is quite well, mostly with the correctness of above 70% in Math, Read, Science, Computer-based problem solving and Digital reading assessment test.

The statistical analyzing result of Math test is comparatively inferior to other subjects, and the situation of computer-based assessment math test presents even worse.

The performance of private school students is more centralized, while public school students disperse more in the test scores.

Although Singaporean students have good reading ability, a part of them get very low scores in the read and digital read assessment. And this situation is especially typical and serious in public school.

In the Math test, students do very well in quantity questions, but have some difficulties in solving practical and geometric problems.

1.2 Perspective Feedback

Specific Reading Enhancement for public school

Schools should offer special education and help to students who are newcomers from not native English speaking countries. And some compulsory language classes and special reading practices can be applied to enhance their reading ability.

More Practical Training in Mathematics Education

In mathematics education, more efforts should be expanded to combine theory with reality, to help students apply what they learn in class into what they meet in life and to form quantitative analyzing ability.

More Computer-Assistant Learning and Teaching

Computer skills, especially statistical and mathematical software, must be involved more in the daily practice and assignments. And students should be trained to recognize and implement visualizations of various mathematical problems.

2. Data Preparation

Originally the dataset 'cogs_sg.csv' given contains 5546 students' testing scores from PISA test. And the test database is composed of 311 questions categorized by 6 different subjects, respectively MATH, READ, SCIE (science), CBAM (computer-based assessment of math), CBAPS (computer-based assessment problem solving) and DRA (digital reading assessment). Students involved in the test came from both public and private secondary schools in Singapore.

2.1 Create subsets by subjects

In this report, different subjects are to be analyzed separately, and there is no blending or mix-up manipulation. Thus I subset the original datatable into 6 different subsets, respectively named as MATH, READ, SCIE, CBAM, CBAPS and DRA, as Figure 1. And each subset should include all the first 9 columns which include the category information of the student like demonstrating the country, school, region etc.

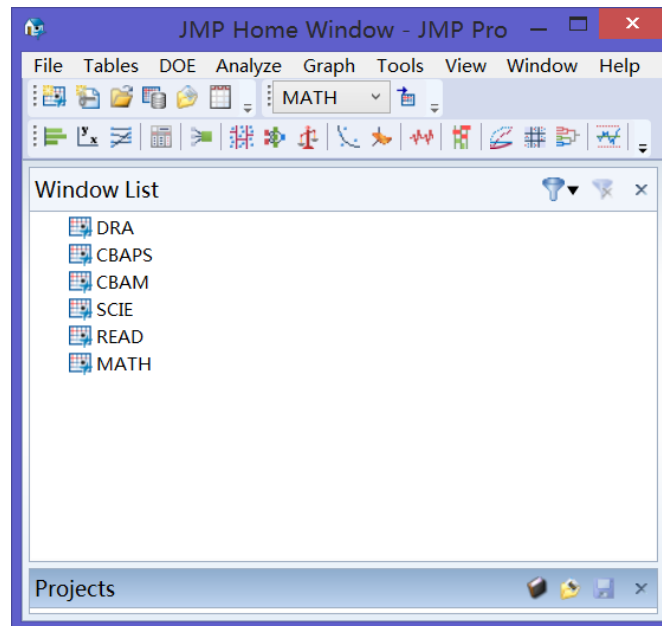


Figure 1

Take MATH subset as an example, as shown in Figure 2. The subset includes all the columns of MATH questions and the information columns.

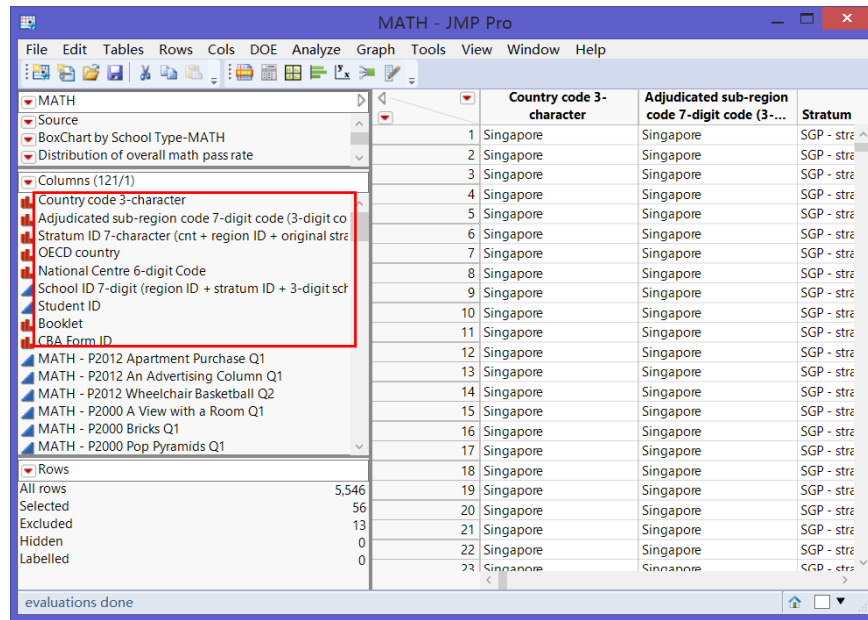


Figure 2

2.2 Recode the attributes

According to codebook of the original dataset, there are 5 different kind of attributes in each cell, respectively standing for different meaning. So firstly I recode these attributes into continuous numbers which can be computed conveniently and directly.

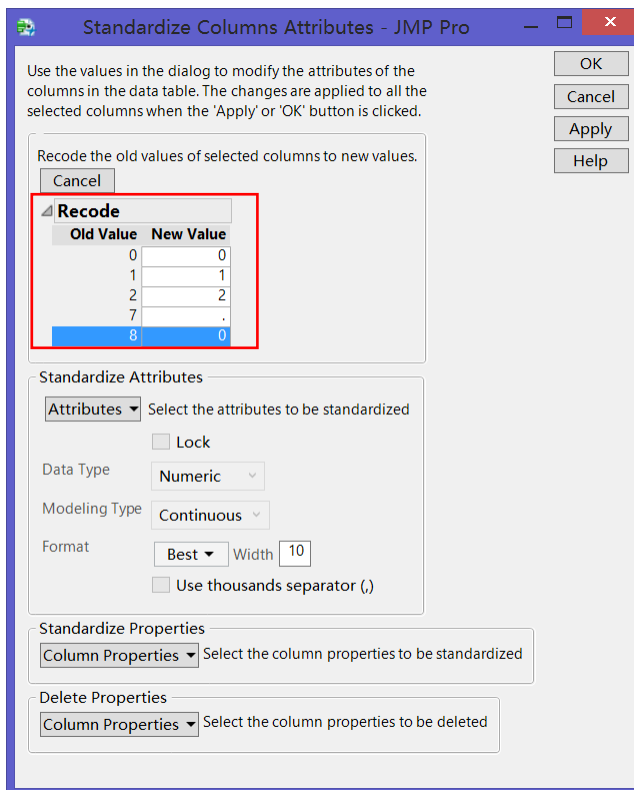


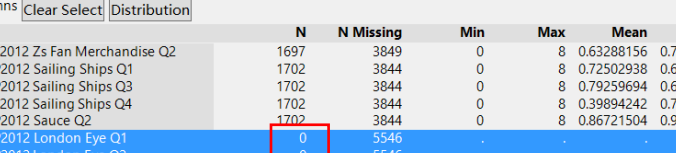
Figure 3

'0', '1' and '2' represent the exact score that students achieved, so no need to be changed. '7' is for No Value, which should not be considered. Then I replace '7' with nothing. Meanwhile '8' is recoded into '0', because students must not get any score from questions they 'Not Reach' as referred in the codebook. (As shown in Figure 3)

One difference between '7' and '8' should be clarified that I count in cells with '8' and recode to 0 because this is when questions appeared in the test but students cannot answer. Nevertheless, questions with '7' never happened in that test to that student, so I choose to ignore them.

2.3 Exclude the all-NA columns and rows within each subset

Before analyzing data, it's necessary to clean the all-NA items within each subset. Take MATH subset as an example in Figure 4 and Figure 5. Check the columns summary and select the columns with zero N which indicate that no students have ever seen these questions. So I exclude these columns.



Summary Statistics

109 Columns

Columns	N	N Missing	Min	Max	Mean	Std Dev
MATH - P2012 Zs Fan Merchandise Q2	1697	3849	0	8	0.63288156	0.77180764
MATH - P2012 Sailing Ships Q1	1702	3844	0	8	0.72502938	0.63008717
MATH - P2012 Sailing Ships Q3	1702	3844	0	8	0.79259694	0.62843961
MATH - P2012 Sailing Ships Q4	1702	3844	0	8	0.39894242	0.70943838
MATH - P2012 Sauce Q2	1702	3844	0	8	0.86721504	0.96864136
MATH - P2012 London Eye Q1	0	5546	-	-	-	-
MATH - P2012 London Eye Q2	0	5546	-	-	-	-
MATH - P2012 Seats In A Theatre Q1	0	5546	-	-	-	-
MATH - P2012 Seats In A Theatre Q2	0	5546	-	-	-	-
MATH - P2012 Racing Q1	0	5546	-	-	-	-
MATH - P2012 Racing Q2	0	5546	-	-	-	-
MATH - P2012 Climbing Mount Fuji Q1	0	5546	-	-	-	-
MATH - P2012 Climbing Mount Fuji Q2	0	5546	-	-	-	-
MATH - P2012 Climbing Mount Fuji Q3	0	5546	-	-	-	-
MATH - P2012 Arches Q1	1697	3849	0	8	0.72304066	0.90979637
MATH - P2012 Arches Q2	1697	3849	0	8	0.25397761	0.99097807
MATH - P2012 Part-Time Work Q1	0	5546	-	-	-	-
MATH - P2012 Part-Time Work Q2	0	5546	-	-	-	-
MATH - P2012 Part-Time Work Q3	0	5546	-	-	-	-
MATH - P2012 Roof Truss Design Q1	1698	3848	0	8	0.88869258	0.67968367

Figure 4

Meanwhile, I check the missing data pattern for all rows. As we can see in Figure 5, there are 13 students who didn't encounter even one MATH question through the test. Thus all these should be regarded as noise and discarded in the MATH subset.

[illegible]

Figure 5

Here comes one key point of data cleaning in this report. All these excluding actions could only be performed within each subset. For example, one student who didn't have math questions is not counted in MATH subset but he could have questions of other subjects, which cannot be ignored. This is why I have to subset the original datatable into 6 different ones.

2.4 Filter out the ‘Double Star’ questions

There are some questions with a total score of 2, which in this report are called ‘Double Star’. ‘Double Star’ of MATH is listed in Figure 6. As referred in the PISA official file, these 2-score questions, whether open questions or multi-choice problems, are more difficult than common one. And students are supposed to spend more time and IQ consumption on these questions. Thus I assume these questions should be weighted more than the 1-score questions.

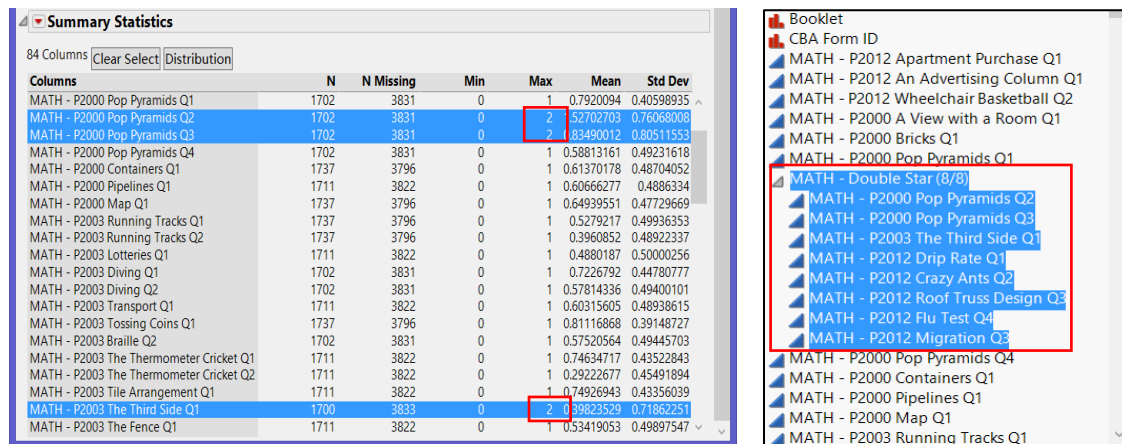


Figure 6

When I compute the number of questions one student have in the test, I count these 2-score questions twice. So if one student has 9 common questions and 1 ‘Double Star’ question, he is assumed to encounter 11 questions in his test. And this rule is used in all 6 subsets.

2.5 Compute the Pass Rate

The item that I use to evaluate one student’s performance in the test is Pass Rate.

$$\text{Pass Rate} = \frac{\text{Sum of Scores}}{\text{Weighted Count of Questions}}$$

I use the rate but not simply sum of scores because, even within one single subset, different students have different number of questions in their test with or without ‘Double-Star’ questions. However as for rate, the corresponding variation can be ignored. Furthermore, Weighted Count of Questions means the hypothetical sum of questions with the assumption that one ‘Double-Star’ question weighted twice as much as common ones.

Take the MATH subset as an example. I first created one new column named as Weighted Count of MATH questions and use Number function to calculate it, as shown in below formulas and Figure 7.

Weighted Count of Questions

$$= \text{Num}(\text{common questions}) + 2 \times \text{Num}(\text{Double} - \text{Star questions})$$

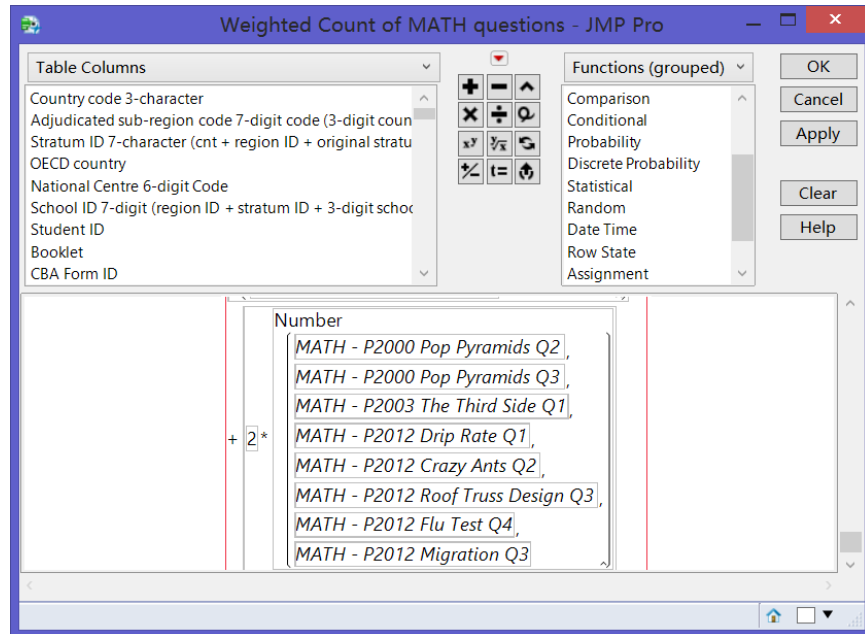


Figure 7

Then the Sum of Score is simply computed by Sum function in JMP formula. Finally, I finish calculating the Pass Rate as Figure 8.

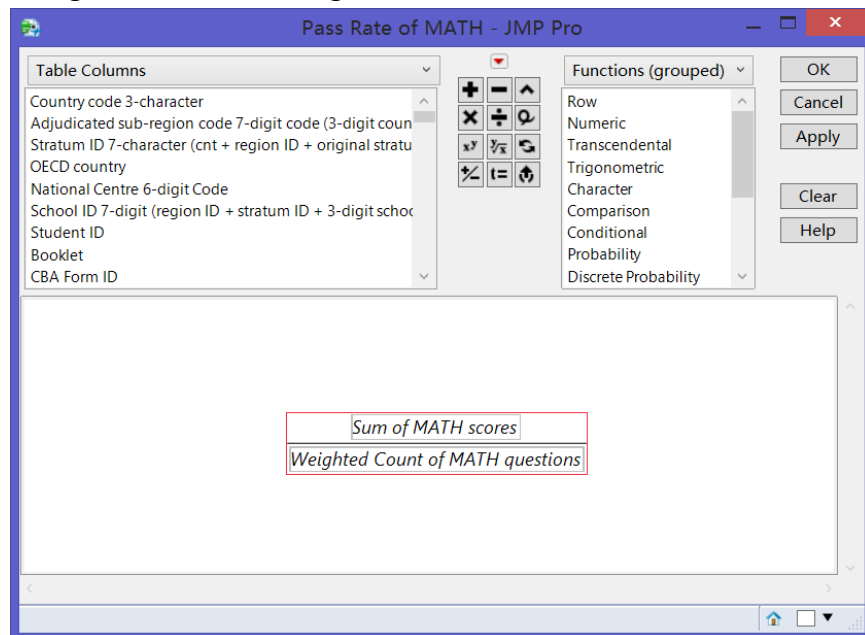


Figure 8

3. Analysis Procedure and Results

3.1 Analyzing General Performance of 6 Subjects

The general performance distribution of 6 subjects is presented as Figure 9. Except the DBAM, all the other subjects present classical left-skewed distribution, and most students' pass rate concentrates on 75%~85%. However, the performance of DBAM distributes more disperse and symmetric, with a larger proportion between 50% and 70%. So I can conclude that generally students perform worst in the DBAM test among all the 6 subjects.

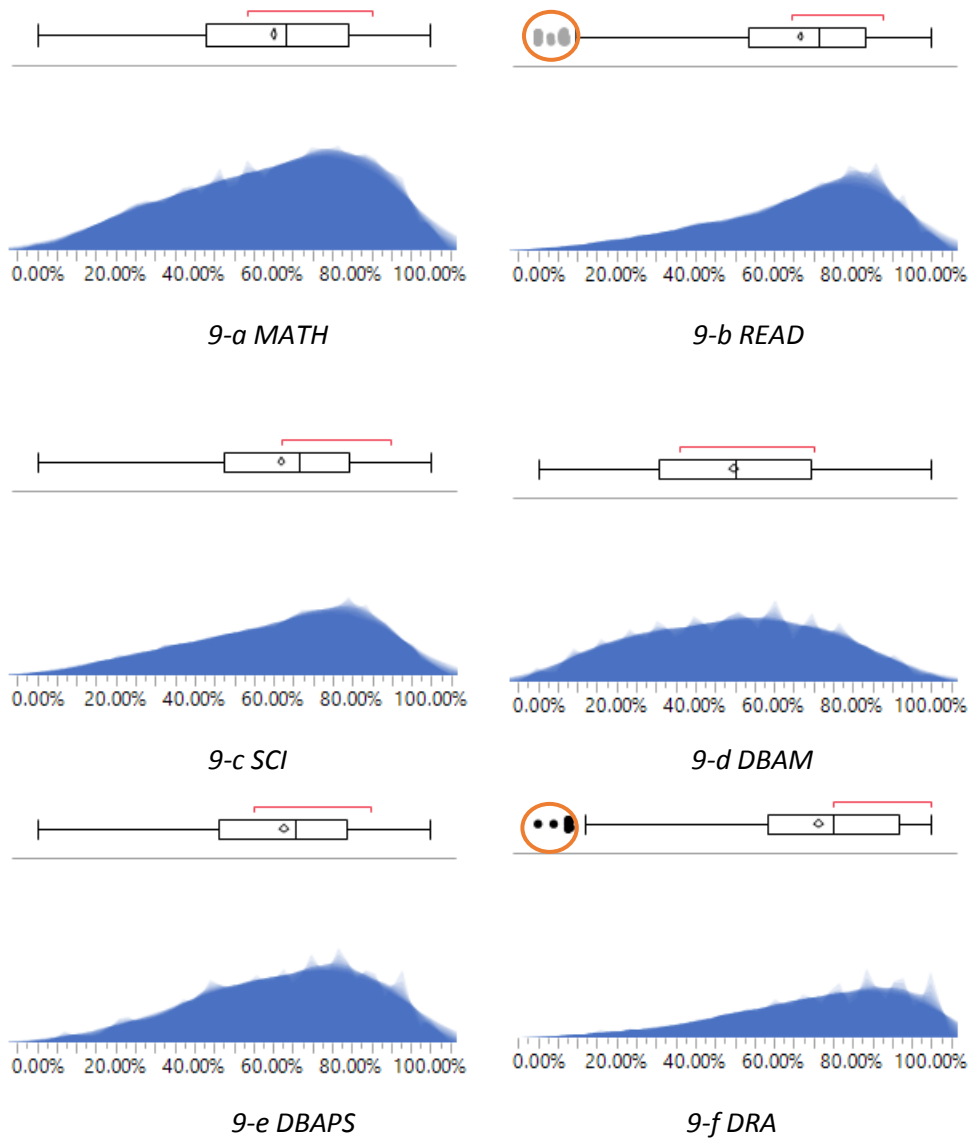


Figure 9 Distribution of general performance of 6 subjects

Furthermore, some obvious outliers can be observed in READ and DRA test with scores as low as less than 20%, as shown in Figure 9-b and 9-f. And this never happens in other

subjects. So there is a part of students who have much lower reading ability than others, which should not be ignored.

The Median and Mean for all subjects are also summarized in Figure 10. For all subjects, median of pass rate is larger than mean, which is also a classical character of left-skewed distribution. Also CRAM and MATH has lower median and mean than other subjects, while CRAM's are much lower.



Figure10

In summary, a quantity of young students in Singapore are able to achieve comparatively high scores in the PISA test. Especially, students perform quite good in READ and DRA test, demonstrating that the overall level of reading capability is higher. However, there exist a small part of students whose reading skills are quite poor and this situation need to be studied more deeply and improved specifically. Moreover, students in Singapore do not get good scores in MATH and CBAM in comparison to the other subjects, while both reflect bad mathematic ability. And students even perform worse in the Computer-based assessment of math. Consequently, it's an important issue to improve the students' ability to use statistical software to solve mathematic problems.

3.2 Comparative Analysis by School Type

Students within the dataset are from two type of schools, public secondary school and private secondary school. The exact number of students from each type of school is shown in Figure 11, which is also summarized from the original dataset in JMP.

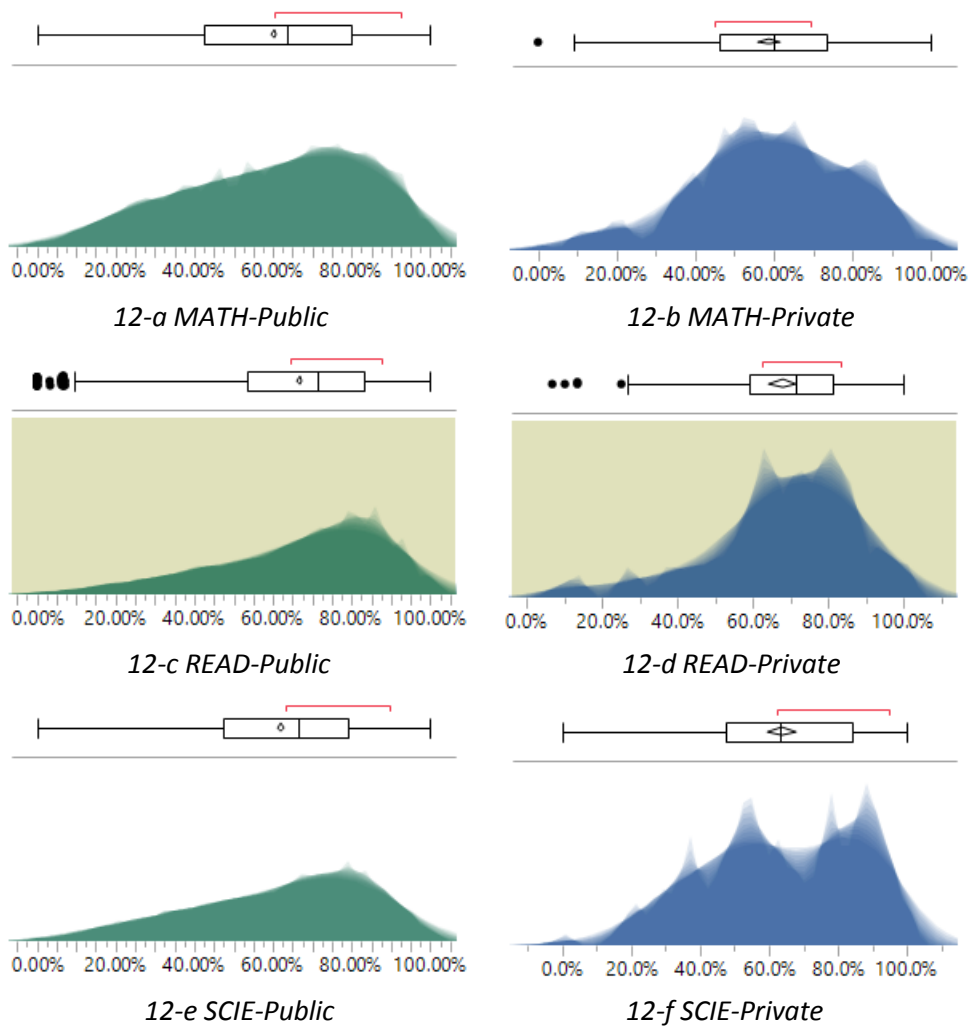
Students from Public school compose a big proportion of the whole dataset. But this corresponds with the realistic situation in Singapore.

Stratum ID 7-character (cnt + region ID + original stratum ID)	N
SGP - stratum 01 : Public Secondary	5369
SGP - stratum 03 : Private Secondary	177

Figure 11

Public schools are supported by government with less tuition fees and are the first choice of most Singaporeans. While the private schools charge quite a great amount of money and the admissions are very limited, so that only few students from high-income family can enter.

The distribution of pass rate for students from different school types are shown in Figure 12. There are quite some differences of the performance between public and private schools within each subset.



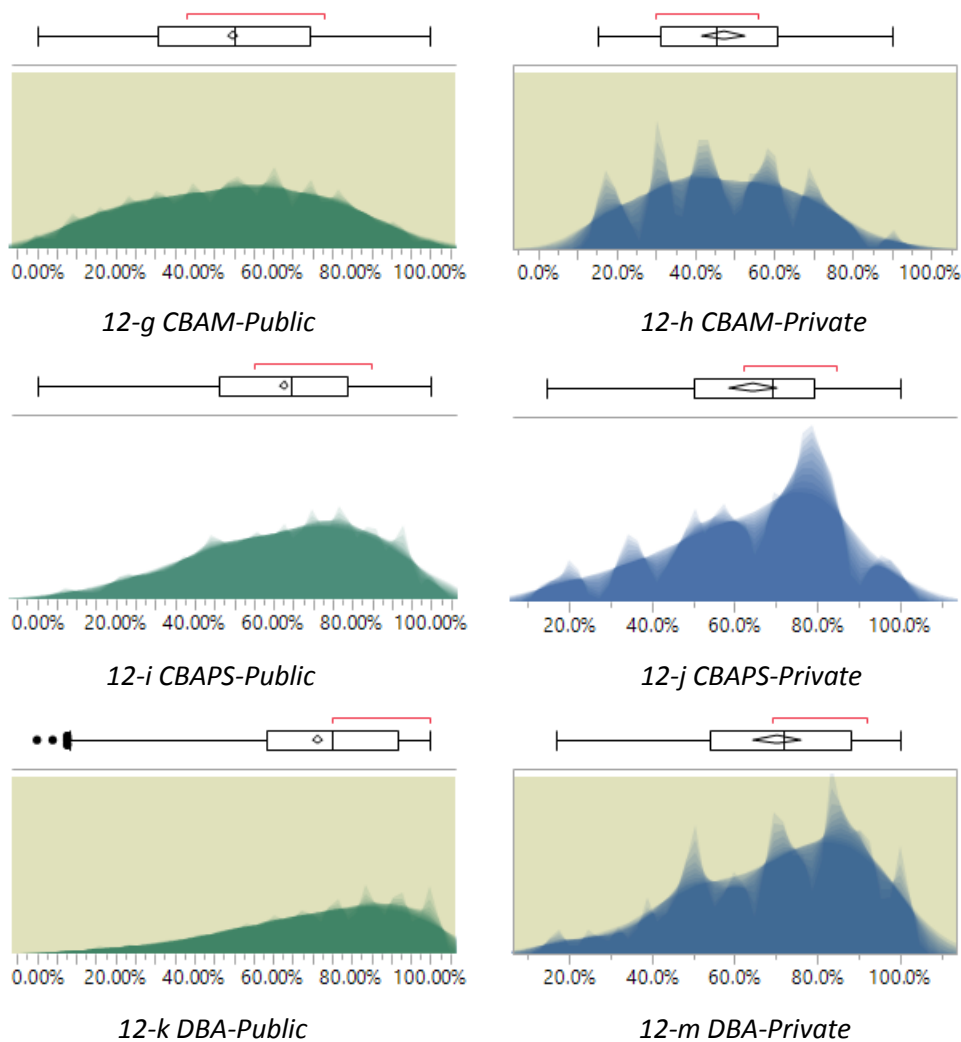


Figure 12 Distribution of 6 subjects by school type

In MATH test, public school students mostly have the pass rate between 70% and 90%, and the distribution is left-skewed. While the performance of private students in MATH trends to be centralized symmetrically between 50% and 70%. Also few private-school students get very low pass rate which compose the outlier in the graph 12-b.

The graphs in 12-c and 12-d, which reflect the difference of READ test performance, have more stories to tell. Firstly a majority of public-school students score better than those from private school. Then also a bigger proportion of students from public school earn very low pass rate than private-school students. Although private school also have this kind of problem, but the proportion is much smaller. This is why a lot of outliers appear in Figure 12-c. The situation of SCIE test is as same as that of MATH, which can be seen in Figure 12-e and 12-f.

In the Computer-based assessment math test, however, the distribution is a little different. Students from both public and private schools do not so well in this test, and the distributions are not so centralized. While the performance of public students are more dispersed. And a great part of the students pass rate falls between 40% and 60%.

In the computer-based assessment problem test, the pass rate of both parts students present classical left-skewed distribution. And most students can get a performance as good as 70%~80%.

Finally in the digital reading assessment, students from both type of schools get very good performance, in which most students' pass rate is even above 80%. And the centralization point of public students is about amazingly 90%. But there are also a few outliers in the distribution of public school, demonstrating that some public school students don't do well in this test. This situation is a little like that in the READ test, as shown in figure 12-k and 12-c.

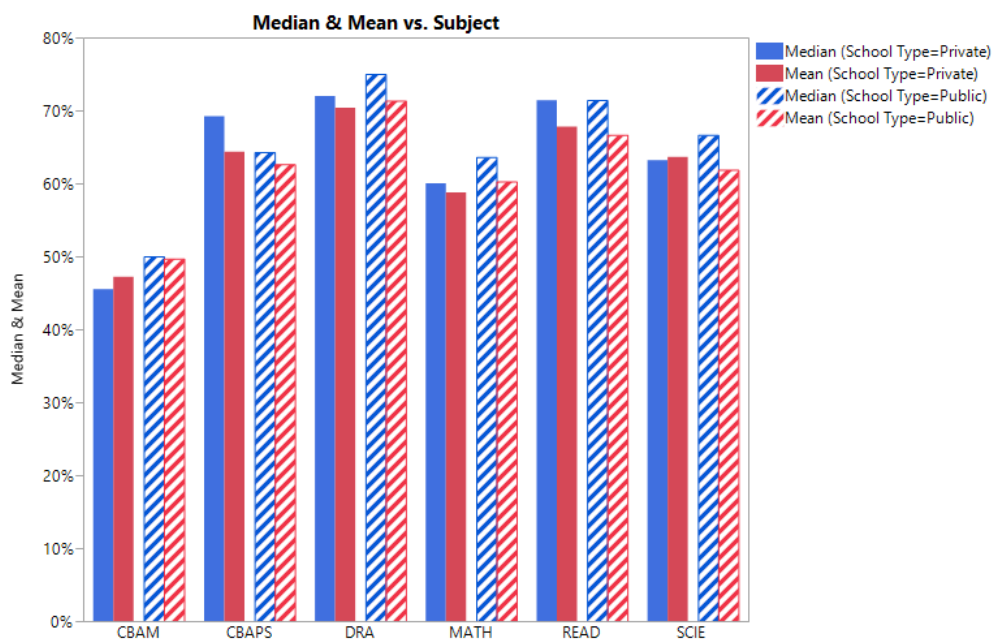


Figure 13

Figure 13 shows the value of mean and median of students from different schools, categorized by 6 subjects. From this view of point, public school students do better than those from private school in CBAM, DRA, MATH, READ and SCIE because the values of median and mean are higher. Only in the CBAPS test, the private school students have better performance.

Generally speaking, for all subjects, private school students' performance are more centralized on a not-so-high level, but almost no one is to be left behind. Students from public schools, on the other hand, have a pass rate more dispersedly distributed. Some

students do really well in the tests but also some other ones are left behind with poor scores. I think this may be because students in private schools are of a small number and the learning resources are more abundant. Thus each student can enjoy enough resources and are not left behind even though they don't tend to work hard.

Especially there are quite a few outliers with too low pass rate existing in the READ test and digital reading assessment test for public school students. So some problems really arise in the education of reading skills right in public school. Maybe we should pay more attention on the low-score students not just the overall performance. And I think this also explains why so many outliers occur in the overall distribution chart of READ and DRA in Figure 9-b and 9-f.

Still one thing need to be mentioned is that both public and private students have bad performance in MATH and computer-based assessment math test. So mathematic capability and statistical software skills are really necessary to be enhanced to the young students.

3.3 Further Analysis of MATH Results

Among all the 311 questions of 6 subjects, MATH questions occupied a proportion of $\frac{1}{3}$. So mathematical ability is the most important item that PISA evaluate. However, students in Singapore have a comparably bad performance in MATH test. And the low pass rate of CBAM also, to some extent, result from bad mathematical ability. So I think it's necessary to have a deep analysis of MATH test results.

According to PISA 2012 Technical Report ANNEX, all the MATH questions can be categorized into four molds as shown in Figure 14. Then I calculate the pass rate of each question molds respectively.

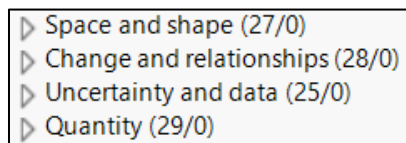


Figure 14

The distribution of MATH pass rate sorted by 4 question molds is shown in Figure 15 and Figure 16. Obviously students' performance varies significantly by different question molds. The easiest question mold for students is Quantity, and the overall pass rate of this mold is about 75%. The questions of uncertainty and data are comparatively not so difficult because the pass rate centralizes between 60~70%.

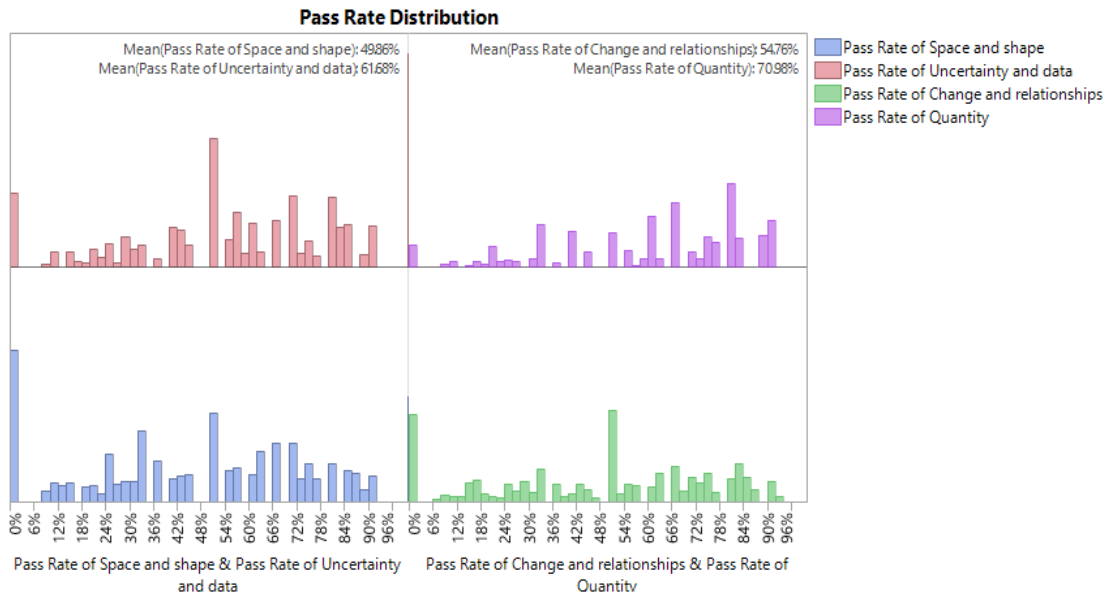


Figure 15

However, when facing Change and relationship questions, students' performance disperses widely. Finally comes the hardest question, it's the Space and shape. Most students can only answer 50% of these questions right. Also some students cannot work out the right answer to even one questions of Space and shape, change and relationships.

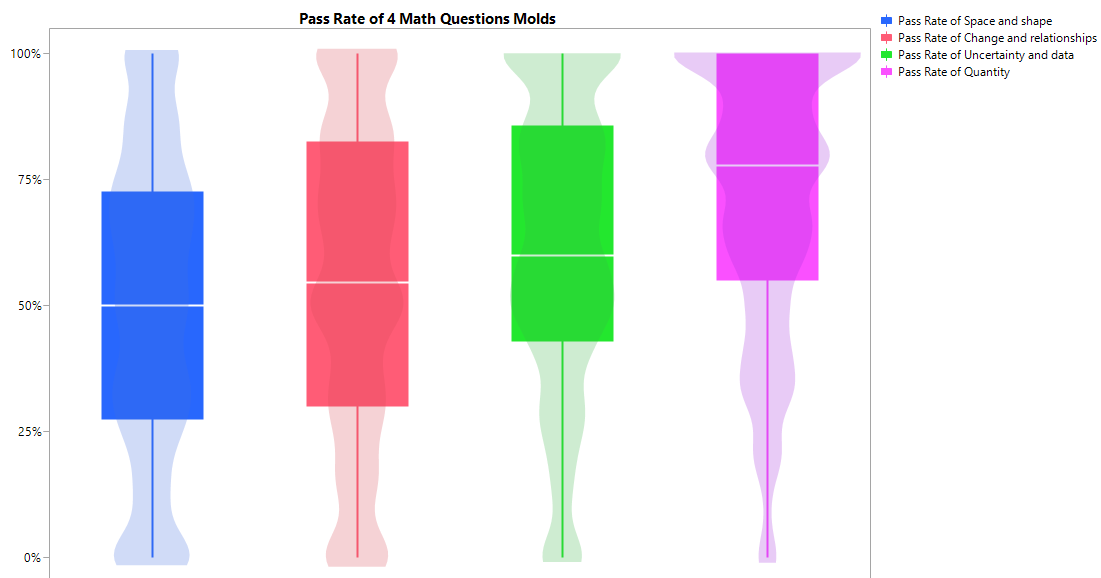


Figure 16

In conclusion, although students don't get nice scores in MATH test, they also have their obvious advantages and disadvantages in different molds of questions. Students in Singapore already do quite well in quantity, but math is not about quantity. What need to be focused on education in math is how to abstract information from shapes and space, something related with geometry, and also understand relationships of different variables.

4. Discussion of Education Improvement

Upon the above analysis results, I have some recommendations about how to improve the education comprehensively and specifically.

4.1 How to Eliminate the Outliers of the READ test?

Among all the tests, students in Singapore have a pretty good overall performance in READ test. However, there are quite a few outliers with very low pass rate that we cannot ignore. And this is especially typical and serious in public school. One reason lead to this problem, as far as I can see, may be that some students come to read in Singapore from other no native English speaking countries. Because all the reading questions are conducted in English and need more English reading ability than other subjects. And students coming from, for example, China are good at math and science but not reading because they have very limited vocabulary.

Thus schools and government should offer special education and help to these students. Some compulsory language classes and special reading practices can be applied to enhance their reading ability.

4.2 Math is not only Formula and Calculation

An understanding of mathematics is central to a young person's preparedness for life in modern society. A growing proportion of problems and situations encountered in daily life, including in professional contexts, require some level of understanding of mathematics, mathematical reasoning and mathematical tools, before they can be fully understood and addressed. Thus it's really necessary and urgent to improve the mathematical ability of young students in Singapore.

The key problem to the mathematics education is the application of mathematical concept and combination with realistic world. Students fail in the math test not because they don't know quantity or calculation but where and how to perform calculations. So students know the formula and can do calculations fast and well, but there are some difficulties for them to abstract information from practical problems and interpret realistic situations. So mathematics teachers are supposed to consider how to combine theory with reality, how to help students apply what they learn in class into what they meet in life and how to form quantitative analyzing ability.

4.3 Digital Teaching and Learning

The widely spread of computers is undoubtedly the major trend of the modern world. And some statistical tools can perform very fast and precise calculations, which will be more efficient for us to solve problems. Also good visualizations of common problems can also be performed in computers. But students in Singapore are not so used to computer-based problems which we can tell by comparing the MATH and CBAM.

So as important as it is, computer skills, especially statistical and mathematical software, must be involved more in the daily practice and assignments.

Reference

1. *PISA 2012 Assessment and Analytical Framework.*
2. *PISA 2012 Technical Report ANNEX A – MAIN SURVEY ITEM POOL CLASSIFICATION.*