

# International Baccalaureate Diploma Program Mathematics Higher Level Internal Assessment

*What is the relationship between High school students' English course scores, TOEFL scores, and SAT Evidence-Based Reading and Writing scores?*

## Table of Contents

1. Introduction .....	1
1.1. Aim of Investigation.....	1
1.2. Rationale.....	1
2. Theory.....	1
2.1. Correlation.....	1
2.2. Regression .....	2
3. Data Collection .....	4
4. Data Analysis.....	4
4.1. Correlation.....	4
4.2. Regression .....	5
5. Conclusion and Evaluation.....	8
5.1. Conclusion.....	8
5.2. Evaluation .....	8
Reference .....	a
Appendix A: Survey .....	b
Appendix B: Original Data .....	c

## 1. Introduction

### 1.1. Aim of Investigation

This investigation aims to find the correlations of high school students' English course scores, TOEFL scores, and their SAT Evidence-Based Reading and Writing scores. If there is a strong correlation between them, then I can create a regression model to use any two factors to predict the third one. This is important for me, because as a high school student studying English as second language, my English proficiency is not as good as my first language, and practicing the external exams such as TOEFL and SAT are hard for me and cost me lots of time and efforts. Therefore, I want to know that to what extent are the practices on the external exams and the English as part of the IB course related to each other, and to what extent can improving practices on one of them help improve the practices on the rest of them.

### 1.2. Rationale

Moreover, using the model, I can estimate a student's SAT score given the student's TOEFL and English course score, or vice versa, estimate the TOEFL score given the SAT and English course score. This helps estimate the student's external exam score without requiring the student to actually know about the exam. This is helpful to me, as I can rationally plan the proper length of time to study one exam when I have only take one exam. The estimation could make suggestions for the time and effort required to practice the exam.

## 2. Theory

### 2.1. Correlation

I use  $e$  to denote a student's English score,  $t$  to denote the student's TOEFL score, and  $s$  to denote the sum of student's SAT Evidence-Based Reading and Writing scores.

Before creating a regression model between two scores, it is important to use a correlation index to evaluate to what extent can the regression model be valid. In order to evaluate the correlation between the two scores, I use Pearson correlation coefficient (Harcet, Heinrichs, Seiler & Skoumal, 2016) to calculate the correlation between any two sets of data among the  $e$ ,  $t$ , and  $s$ . I use  $\rho_{X,Y}$  to represent the correlation between  $X$  and  $Y$ , where  $X$  and  $Y$  are any two sets of data.

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sqrt{\sigma_X^2 \sigma_Y^2}} \quad (1)$$

In the equation,  $\sigma_{X,Y}$  is the covariance between the two sets of data  $X$ ,  $Y$  (Harcet et al., 2016),

$$\sigma_{X,Y} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n} \quad (2)$$

and  $\sigma_X^2$  is the variance of dataset  $X$ , and similarly,  $\sigma_Y^2$  is the variance of dataset  $Y$  (Harcet et al., 2016).

$$\sigma_X^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} \quad (3.1)$$

$$\sigma_Y^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n} \quad (3.2)$$

How can I represent the correlations between any two sets of numbers? I can construct a correlation matrix to illustrate the correlations between any two sets of data.

$$Corr = \begin{bmatrix} \rho_{e,e} & \rho_{e,t} & \rho_{e,s} \\ \rho_{t,e} & \rho_{t,t} & \rho_{t,s} \\ \rho_{s,e} & \rho_{s,t} & \rho_{s,s} \end{bmatrix} \quad (4)$$

In the equation, there is an important property, that the variables in the main diagonal, viz.  $\rho_{e,e}$ ,  $\rho_{t,t}$ ,  $\rho_{s,s}$ , all have the value of 1, because the two variables are the same, as shown in equations (5).

$$\rho_{X,X} = \frac{\sigma_{X,X}}{\sqrt{\sigma_X^2 \sigma_X^2}} \quad (5.1)$$

$$= \frac{\sigma_{X,X}}{\sigma_X} \quad (5.2)$$

$$= \frac{\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})}{\frac{n}{\sum_{i=1}^n (X_i - \bar{X})^2}} \quad (5.3)$$

$$= \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad (5.4)$$

$$= 1 \quad (5.5)$$

Therefore, the correlation matrix could be directly written as equation (6).

$$Corr = \begin{bmatrix} 1 & \rho_{e,t} & \rho_{e,s} \\ \rho_{t,e} & 1 & \rho_{t,s} \\ \rho_{s,e} & \rho_{s,t} & 1 \end{bmatrix} \quad (6)$$

Quite interesting, huh!

## 2.2. Regression

There are multiple ways to produce a hypothesis equation using two variables, including linear regressions and polynomial regressions. For example, if I use linear regression, the hypothesis equation could be written as equation (7). The hypothesized value is represented with  $h_{x,y}$ , given that  $x$  and  $y$  are two known variables.

$$h_{x,y} = \theta_0 + \theta_1 x + \theta_2 y \quad (7)$$

In the equation,  $\theta_i$  are the parameters for the regression.

In order to optimize the parameters  $\theta_i$  for the hypothesis equations, I use the least square method, minimizing the difference between the hypothesized value and the real value.

This is actually similar to machine learning in the process of minimizing the squared difference! Wow! I use  $J$  to represent the sum of the squared difference between the hypothesized value and the real value (Ng, n.d.).

$$J = \sum_{i=0}^n (h_{x,y} - z_i)^2 \quad (8)$$

In equation (8),  $z_i$  is the real value and  $h_{x,y}$  represents the hypothesized value. In the equation,  $x_i$ ,  $y_i$ , and  $z_i$  are all given variables and the objective is to adjust the parameters  $\theta_i$  in the following equation to minimize  $J$ .

Combining equation (7) and equation (8), I get equation (9) to make the further steps easier.

$$J = \sum_{i=0}^n (\theta_0 + \theta_1 x_i + \theta_2 y_i - z_i)^2 \quad (9)$$

In order to minimize  $J$ , I use the gradient descent method, iterating the parameters  $\theta_i$ , subtracting the partial derivative times a constant (learning rate) simultaneously for all parameters  $\theta_i$  (Ng, n.d.). This is actually the same process with machine learning, and I feel proud for the depth of exploration I'm taking.

$$\theta_{0,new} = \theta_0 - \alpha \frac{\partial J}{\partial \theta_0} \quad (10.1)$$

$$\theta_{1,new} = \theta_1 - \alpha \frac{\partial J}{\partial \theta_1} \quad (10.2)$$

$$\theta_{2,new} = \theta_2 - \alpha \frac{\partial J}{\partial \theta_2} \quad (10.3)$$

But what does the 'partial' mean? How do I calculate the partial derivatives? Actually it is not hard, similar with all the mathematics, if only I try to learn. I realized that to calculate partial derivatives is just to calculate the derivative as if the other variables is constant. Therefore, I quickly calculate the partial derivative of the theta in equation (11).

$$\frac{\partial J}{\partial \theta_0} = 2 \sum_{i=0}^n (\theta_0 + \theta_1 x_i + \theta_2 y_i - z_i) \quad (11.1)$$

$$\frac{\partial J}{\partial \theta_1} = 2x_i \sum_{i=0}^n (\theta_0 + \theta_1 x_i + \theta_2 y_i - z_i) \quad (11.2)$$

$$\frac{\partial J}{\partial \theta_2} = 2y_i \sum_{i=0}^n (\theta_0 + \theta_1 x_i + \theta_2 y_i - z_i) \quad (11.3)$$

Substituting the partial derivative in equations (10) with equations (11), I get the final iteration equations.

$$\theta_{0,new} = \theta_0 - 2\alpha \sum_{i=0}^n (\theta_0 + \theta_1 x_i + \theta_2 y_i - z_i) \quad (12.1)$$

$$\theta_{1,new} = \theta_1 - 2\alpha x_i \sum_{i=0}^n (\theta_0 + \theta_1 x_i + \theta_2 y_i - z_i) \quad (12.2)$$

$$\theta_{2,new} = \theta_2 - 2\alpha y_i \sum_{i=0}^n (\theta_0 + \theta_1 x_i + \theta_2 y_i - z_i) \quad (12.3)$$

Following these equations, I will obtain the parameter values in the hypothesis equation, therefore predict the third score.

### 3. Data Collection

Where does the data come from? Since these kind of data is highly confidential and may influence the interpersonal relationships between students if leaked, I use an anonymous online survey to collect the data. You can check Appendix A to see the survey.

In the survey, question 1 & 2 is actually not the information I need, but I believe that these choices could make the survey seems more formal, and therefore get a more reliable result. Moreover, the information may help me better understand the sources of my data. Actually I carefully designed questions 5 & 6 to be separated, because I think some students may carelessly input the total SAT score in the entry at question five. Thus, question 6 may help them realize the wrong information they entered. In the end, I added question 7 to improve my data quality, as students may feel guilty to tick the button if they intentionally input the wrong data. I feel my survey deliberately designed and would yield successful results.

### 4. Data analysis

#### 4.1. Correlation

To my surprise, I collected 65 data, but only 29 data can be used after tidying up! It's frustrating, but these data are still enough for my investigation. There are three features of data: English course score, denoted as  $e$ ; TOEFL score, denoted as  $t$ ; and SAT Evidence-Based Reading and Writing, denoted as  $s$ . I plot the relationship between each two of the three sets of data in Figure 1 to Figure 3.

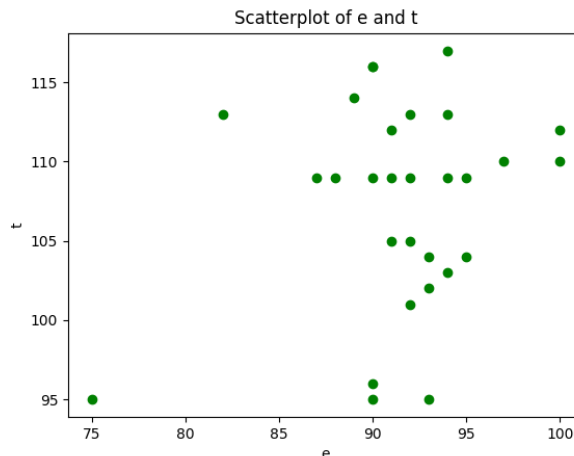


Figure 1

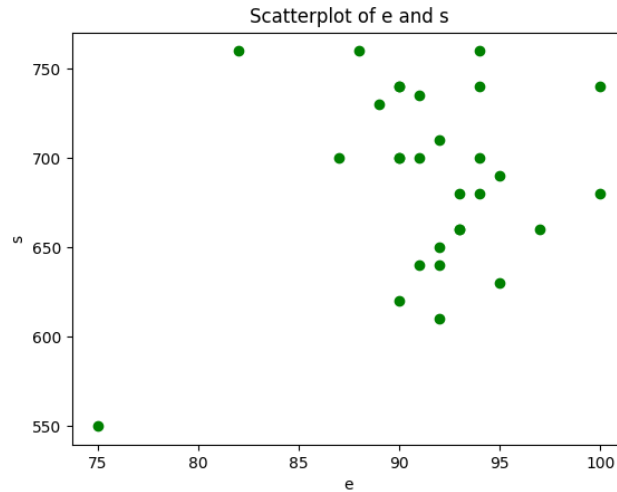


Figure 2

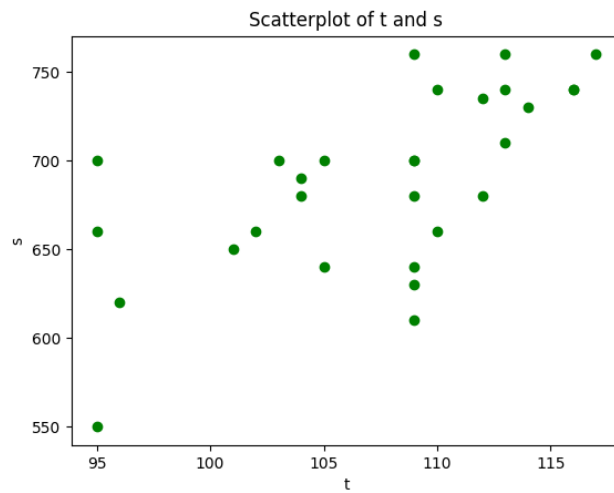


Figure 3

Looking at the graph first, there is small relationship between  $e$  and  $t$ , and  $e$  and  $s$ , but there is clear relationship between  $t$  and  $s$ . Calculate the correlation as suggested in theory, I get the correlation matrix as equation (13).

$$\begin{bmatrix} 1 & 0.225 & 0.183 \\ 0.225 & 1 & 0.638 \\ 0.183 & 0.638 & 1 \end{bmatrix} \quad (13)$$

The correlation matches the graph.

#### 4.2. Regression

As there is not a strong relationship between  $e$  and  $t$ , and  $e$  and  $s$ , I decide to change the plan a little bit, to regress using  $t$  and  $s$  directly.

There are two ways of regression, one is using  $t$  to predict  $s$ . As there are only two variables in the regression, the theory is updated as equation (14) to (16).

$$h_t = \theta_0 + \theta_1 t \quad (14)$$

$$J = \sum_{i=0}^n (h_t - s_i)^2 \quad (15)$$

$$\theta_{0,new} = \theta_0 - 2\alpha \sum_{i=0}^n (\theta_0 + \theta_1 t_i - s_i) \quad (16.1)$$

$$\theta_{1,new} = \theta_1 - 2\alpha t_i \sum_{i=0}^n (\theta_0 + \theta_1 t_i - s_i) \quad (16.2)$$

After iteration according to equations (14), I obtain the result in equation (17).

$$\theta_0 = 145 \quad (17.1)$$

$$\theta_1 = 5.06 \quad (17.2)$$

Therefore,

$$s = 145 + 5.06t \quad (17.3)$$

I plot the graph of the scattered points of the two variables and the line of the hypothesized equation in Figure 4.

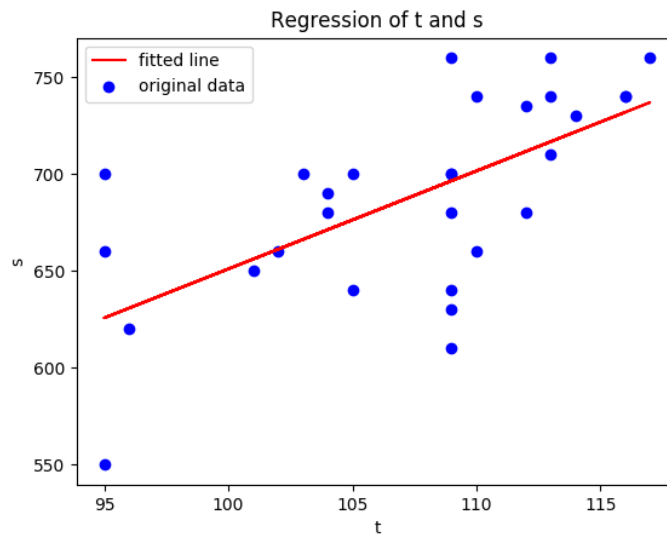


Figure 4

Similarly, the other way of prediction is using  $s$  to predict  $t$ , shown in equation (18).

$$t = 51.9 + 0.0806s \quad (18)$$

I also plot the graph of the scattered points and the line of the hypothesized equation in Figure 5.

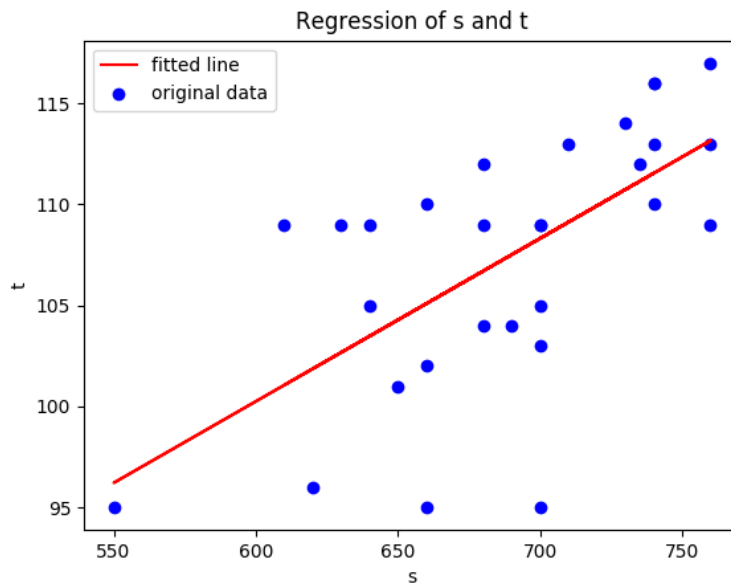


Figure 5

Hence I can predict a student's SAT score given his TOEFL score, and vice versa. Great! For example, if a student has a TOEFL score of 110, the student's SAT Evidence-Based Reading and Writing score is predicted in equation (19).

$$s = 145 + 5.06t = 145 + 5.06 \times 110 = 702 \quad (19)$$

Similarly, if a student has a SAT Evidence-Based Reading and Writing score of 630, his TOEFL score is predicted in equation (20.1).

$$t = 51.9 + 0.0806s = 51.9 + 0.0806 \times 630 = 103 \quad (20.1)$$

Not bad!

However, note that the if I put the TOEFL score of 103 back to the first model, the SAT score would be predicted as follows in equation (20.2).

$$s = 145 + 5.06t = 145 + 5.06 \times 103 = 666 \quad (20.2)$$

This is different as the original SAT score input of 630. At first time I thought my program went wrong, but after several times of checking, I can be sure that it didn't. I think, and think, and think. Suddenly, I realized that this is due to the optimization goal difference! While the first model minimizes the least squared error between the predicted SAT score and the real SAT score, given the students' TOEFL score, the second model minimizes the least squared error between the predicted TOEFL score and the real TOEFL score, given the students' SAT score. However, they are all accurate regressions, and the hypothesis equation all represents the best predicted value.



## 5. Conclusion and Evaluation

### 5.1. Conclusion

There is little relationship between students' English course scores and the other two scores, viz. TOEFL scores and the SAT Evidence-Based Reading and Writing scores. However, there is a relationship between students' TOEFL scores and the SAT Evidence-Based Reading and Writing scores, suggesting the feasibility to create a regression model between these two scores. I do create a linear regression model to predict one of the related scores given the other, and it clearly shows the linear relationship between these two scores.

However, given that the relationship is not very strong, the accuracy of the prediction is pretty limited. It would be definitely better if I collect more dimensions of data and analysis the relationships between them. By this method, I could make the regression model far more accurate and yield better results.

Moreover, given that I collected only 28 effective data, and all of which are from our high school, the model would be biased. I don't have any data from students outside our school to represent their situations, and such a low proportion of the data could not necessarily represent the entire situation in our school. I wish that I could collect more data, but the survey distribution is pretty limited. If I continue to conduct my research, I shall try to apply or raise some funds as a stimuli to encourage people to come and fill my survey.

### 5.2. Evaluation

I use Pearson correlation coefficient rather than Spearman's rank correlation coefficient to calculate the correlation between any two sets of data. This is because Pearson correlation calculates the linear relationships between two variables, which is what measures the rationality of creating a linear regression model. However, Spearman's rank correlation represents the extent of monotonic relationship between two sets of data, and doesn't necessarily represent the rationality of creating a linear regression model.

My current research can only show the correlation between a student's English course score, TOEFL score, and SAT Evidence-Based Reading and Writing score, but cannot suggest the causation between one another. However, it is still reasonable to create a linear regression model to predict the possible performance of the untaken exam, according to the correlation between them.

I need to convert the ACT scores to SAT scores because some of the students in my sample choose to take ACT test. I didn't expect this. There is a relatively trustworthy conversion table on the internet and I manually convert their ACT score to SAT score. As ACT has a similar structure, I convert the ACT score in the related two blanks, which is assumed the ACT English and Reading scores, and divide them by two to get the corresponding SAT Evidence-Based Reading and Writing score. This may cause some error of my data.

There are some careless students who fill my survey extra fast and misinterpreted the meaning of my question. There are also some students who intentionally want to create some noise in my data. Though part of them could be filtered by the final question, some

of them may still be bold enough to deliberately add outliers in my data. This may cause some error in my model.


There are students who strategically take TOEFL test for the first time, and concentrate their effort to conquer SAT, taking it for several times, and finally take another TOEFL test. This would cause significant error in their TOEFL scores as their former TOEFL score couldn't represent their current performance on this test. However, this trivial situation is too complex to be verified and excluded in my model and thus cause error. In the other hand, there are also students who take SAT for the first time and concentrate their effort to conquer TOEFL test, whose data balanced the extreme values from the former situation.

There are multiple ways of regression, but I simply use linear regression with one variable, to construct my model. The most important reason is that there are no significant correlations between two sets of data rather than TOEFL score and the SAT Evidence-Based Reading and Writing score. Moreover, I don't have a huge amount of data to let me construct a more complex model without overfitting.

## Reference

- Fracchia, K. (2016, January 31). *ACT to new SAT to old SAT score conversion chart*. Retrieved from <https://magoosh.com/hs/act/act-scores/2016/act-to-new-sat-to-old-sat-score-conversion-chart/>
- Harcet, J., Heinrichs, L., Seiler, P., Skoumal M. (2016). Covariance. *Mathematics higher level: statistics*. Oxford, England: Oxford University Press.
- Ng, A. (n.d.). *Gradient descent for linear regression* [Video File]. Retrieved from <https://www.coursera.org/learn/machine-learning/lecture/kCvQc/gradient-descent-for-linear-regression>.

## Appendix A: Survey



### Help Diego's Math IA

This survey takes about 3 minutes. Thank you for your time.

S2C6 Diego is currently doing his math internal assessment, aiming to find relationships between students' English grade, TOEFL grade and SAT grade in order to help students predict their score. You can contribute to this research by providing your data anonymously.

---

1. Your grade \*

Senior 1  
 Senior 2  
 Senior 3

---

2. Your curriculum \*

IGCSE  
 A-level  
 IB  
 AP

---

3. Your current English percentage grade \*

---

4. Your TOEFL score (if taken)

---

5. Your SAT reading score (if taken)

---

6. Your SAT writing & language score (if taken)

---

7. Thank you for your contribution.

I'm sure my data is real.

---

---

问卷星 提供技术支持

## Appendix B: Original Data

<i>number</i>	<i>1. Your grade</i>	<i>2. Your curriculum</i>	<i>3. Your current English percentage grade</i>	<i>4. Your TOEFL score (If taken)</i>	<i>5. Your SAT reading score (If taken)</i>	<i>6. Your SAT writing &amp; language score (If taken)</i>	<i>7. Thank you for your contribution.</i>
1	Senior 2	IB	95	109	300	330	I'm sure my data is real.
2	Senior 2	IB	60	60	60	60	I'm sure my data is real.
3	Senior 2	IB	100	112	320	360	I'm sure my data is real.
4	Senior 2	IB	94	109	330	350	(空)
5	Senior 3	IB	93	95	1450	656	I'm sure my data is real.
6	Senior 2	IB	92	109	1410	610	(空)
7	Senior 2	IB	92	105	300	340	I'm sure my data is real.
8	Senior 2	IB	85	104	(空)	(空)	I'm sure my data is real.
9	Senior 2	IB	7	101	650	700	I'm sure my data is real.
10	Senior 2	IB	94	113	740	750	I'm sure my data is real.
11	Senior 2	IB	90	116	740	800	I'm sure my data is real.
12	Senior 2	IB	85%	105	32	(空)	(空)
13	Senior 2	IB	100	无	无	无	I'm sure my data is real.
14	Senior 2	IB	90	96	290	330	I'm sure my data is real.
15	Senior 1	AP	B	(空)	(空)	(空)	I'm sure my data is real.
16	Senior 2	IB	100	110	370	370	I'm sure my data is real.
17	Senior 2	IB	high enough to nail you	high enough to nail you	high enough to nail you	high enough to nail you	I'm sure my data is real.
18	Senior 2	IB	94	103	360	340	I'm sure my data is real.
19	Senior 1	IB	89	106	(空)	(空)	I'm sure my data is real.
20	Senior 1	A-level	89	108	(空)	(空)	I'm sure my data is real.
21	Senior 1	A-level	91%	(空)	(空)	(空)	I'm sure my data is real.
22	Senior 1	AP	95	94	(空)	(空)	I'm sure my data is real.
23	Senior 3	IB	87	109	320	380	I'm sure my data is real.
24	Senior 1	A-level	81	102	(空)	(空)	I'm sure my data is real.
25	Senior 2	IB	90	116	370	370	I'm sure my data is real.
26	Senior 2	A-level	93	102	660	no	I'm sure my data is real.
27	Senior 2	IB	94	117	360	400	I'm sure my data is real.

28	Senior 2	IB	88	109	32	35	I'm sure my data is real.
29	Senior 2	IB	89	106	1500	15	I'm sure my data is real.
30	Senior 2	AP	100	109	1510	(空)	I'm sure my data is real.
31	Senior 2	A-level	7	102	ACT28	(空)	I'm sure my data is real.
32	Senior 2	IB	88	108	330	380	I'm sure my data is real.
33	Senior 1	IGCSE	90	95	700	6	I'm sure my data is real.
34	Senior 2	IB	93	104	680	12	I'm sure my data is real.
35	Senior 2	IB	89	114	I forgot	Total of 730.	I'm sure my data is real.
36	Senior 2	IB	91	112	act32	act 8	I'm sure my data is real.
37	Senior 2	IB	91	105	1480	700	I'm sure my data is real.
38	Senior 2	IB	89	110	(空)	(空)	I'm sure my data is real.
39	Senior 2	IB	82	113	380	380	I'm sure my data is real.
40	Senior 1	AP	92	112	(空)	(空)	I'm sure my data is real.
41	Senior 2	A-level	97	(空)	(空)	(空)	I'm sure my data is real.
42	Senior 2	AP	95	(空)	(空)	(空)	I'm sure my data is real.
43	Senior 2	AP	98	110	(空)	(空)	I'm sure my data is real.
44	Senior 2	AP	98	105	(空)	(空)	I'm sure my data is real.
45	Senior 1	IB	93%	115	(空)	(空)	I'm sure my data is real.
46	Senior 3	IB	7	(空)	(空)	(空)	(空)
47	Senior 2	AP	90	113	0	0	I'm sure my data is real.
48	Senior 1	A-level	91	(空)	(空)	(空)	I'm sure my data is real.
49	Senior 1	AP	93	95	(空)	(空)	I'm sure my data is real.
50	Senior 2	A-level	90	(空)	(空)	(空)	(空)
51	Senior 1	AP	88	confidential	No	No	I'm sure my data is real.
52	Senior 1	AP	93	102	(空)	(空)	I'm sure my data is real.
53	Senior 1	IGCSE	30	12	(空)	(空)	(空)
54	Senior 2	A-level	75%	95	250	300	I'm sure my data is real.
55	Senior 1	IB	90	109	330	370	I'm sure my data is real.
56	Senior 3	IB	92	113	370	340	I'm sure my data is real.
57	Senior 1	IB	96	111	(空)	(空)	I'm sure my data is real.
58	Senior 3	IB	91	109	290	350	I'm sure my data is real.

59	Senior 2	A-level	97%	110	330	330	I'm sure my data is real.
60	Senior 2	AP	95%	104	340	350	I'm sure my data is real.
61	Senior 2	A-level	95	111	(空)	(空)	I'm sure my data is real.
62	Senior 3	IB	92	112	1490	(空)	I'm sure my data is real.
63	Senior 1	IGCSE	80	87	(空)	(空)	I'm sure my data is real.
64	Senior 2	AP	98	93	1420	4 5 6	I'm sure my data is real.
65	Senior 2	AP	93	102	1430	444	I'm sure my data is real.