VIETNAM JOURNAL OF EDUCATION, 2024
Volume 8, Issue 1, 1-17
https://doi.org/10.52296/vje.2024.334

ORIGINAL ARTICLE

A Statistical Analysis of Vietnamese High School English Test Score Data

Thai Ton Duong Duong[1], Uyen To Huynh[1,2], Nabendu Pal[3,4,+]

[1]Vietnam National University, Vietnam;
[2]University of Economics and Law, Vietnam;
[3]Faculty of Mathematics and Statistics, Ton Duc Thang University, Vietnam;
[4]Department of Mathematics, University of Louisiana at Lafayette, USA
[+]*Corresponding author ● Email: nabendu.pal@tdtu.edu.vn*

Article history
Received: 07 October, 2023
Accepted: 28 December, 2023
Published: 28 February, 2024

Keywords
Analysis of variance (ANOVA), national high school examination, Vietnamese education, provincial mean test score, provincial median test score

ABSTRACT
This statistical study, perhaps the first of its kind in Vietnamese academia, has tried to understand the patterns in high school students' pre-college national exam test scores which hold tremendous significance in students' college or university admissions. English belongs to one of the three mandatory subjects (the other two being Mathematics and Literature) where students must be tested as part of their overall evaluation. However, since English is a foreign language for the learners, many high school students, especially those from rural areas, struggle to attain an acceptable level of proficiency. Meanwhile, a good proportion of students, particularly those in urban areas and with their own substantial financial resources, do attend extra tutorial classes to gain an edge in English proficiency. In this study we have analyzed the provincial mean and median test scores over a period of three years (2019, 2020 and 2021). It has been noted that a province's aggregate performance depends heavily on - (i) the province's per capita gross domestic product (PCGDP), which measures a province's economic prosperity; and (ii) the average number of students per high school (ANSHS) which indicates the overall competitive environment the students encounter which helps them sharpen their language proficiency.

1. INTRODUCTION

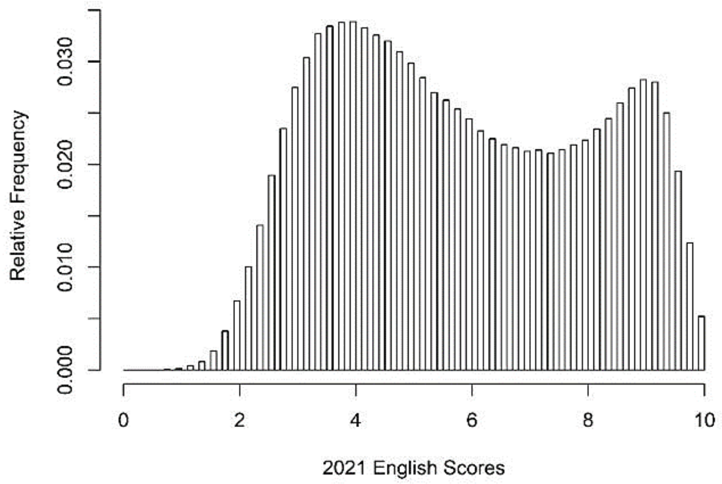
Some recent facts

Every year, as observed over the last five years, approximately a million Vietnamese students sit the National High School Graduation Examination (NHSGE) which has a strong implication for their college and/or university admissions. NHSGE takes place in late summer, and each student must take tests in the mandatory trio - mathematics, literature and a foreign language. In the foreign language category even though the students have other options like Chinese, French, German, Japanese and Russian, it is estimated that about 99% of the students end up learning English, and subsequently take the English test during NHSGE, performance in which then becomes a part of the national dialogue, causes plenty of heartburn and soul-searching.

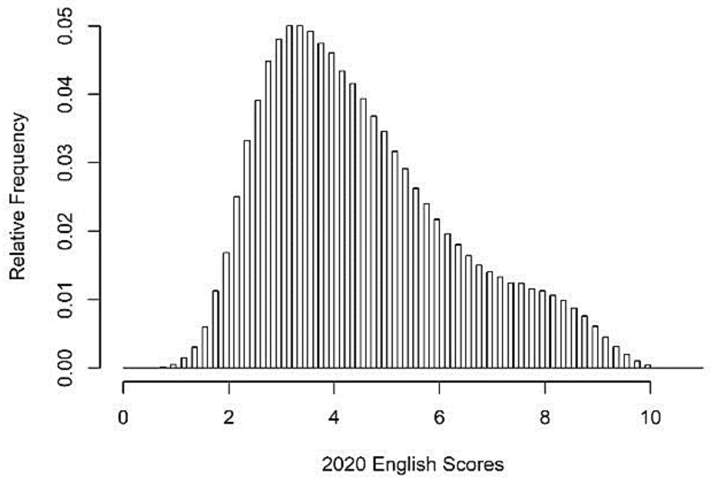
On a scale from 0 to 10, the national average English score in 2020 fell to 4.58, below the halfway mark, and thus causing a great deal of consternation among the students, parents as well as the educators nationwide. The 2020 results had even more depressing information to ponder over. For example, a staggering 63.1% of the examinees scored below the average score (i.e., 4.38), and a little more than 50% of the students made an average score of 3.4. Even though the 2020 average score of 4.58 was a tad above the 2019 average score, still it was the worst compared to other subject areas; such as - the average History score was 5.19, the average Math score was 6.68, the average Literature score was 6.62, and the average Civic score was an impressive 8.14.

While the 2020 average score (of 4.58) may look depressing, it also belies another profound fact, that is, the ‘urban - rural divide’. The top four areas (or provinces, in terms of highest score) were the heavily urban (and/or industrial) ones - Ho Chi Minh City (HCMC), Binh Duong, Ba Ria - Vung Tau, and Hanoi, with HCMC having the highest average score of 5.85. It is not at all surprising that the provinces which are mostly agrarian and less industrialized, such as the northern mountainous ones, like Hoa Binh, Ha Giang and Son La, achieved the lowest scores. The glaring regional differences were also observed in the most recent 2021 English test scores.

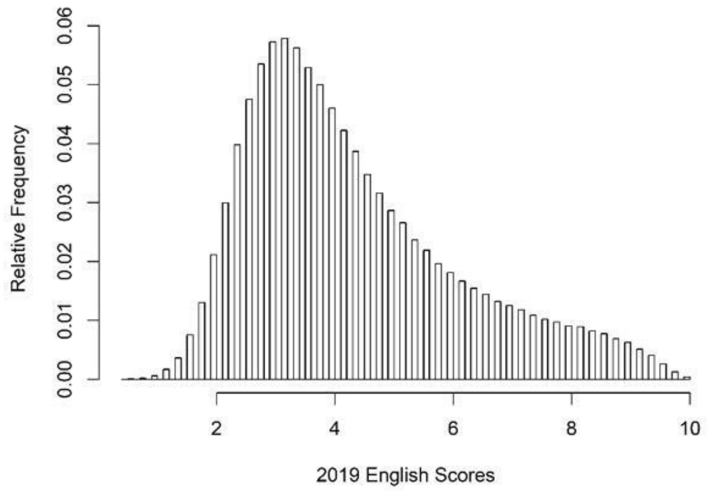
The English score distribution (see the following Figure 1 (a)) showed two distinct peaks, one around 3.80, and the other around 9.00. The lower peak perhaps represents the predominantly rural students with less resources, and the higher peak represents the more resourceful and advantageous students whose families could afford to send them to better schools or provide them with extra private tutoring which happens mostly in the urban areas. The following three figures (Figure 1 (a) - (c)) depict the overall picture of the national English test performance through the relative frequency histograms of the scores for the three most recent years.



(a)



(b)



(c)

Figure 1. Relative frequency histogram of the English test scores; (a) 2021, (b) 2020, and (c) 2019

The ground realities

English is now a compulsory subject taught from the third grade in Vietnam. Before 2010, students were introduced to this language from the sixth grade. Obviously, since 2010, some improvements have taken place, however minimal they might be. In 2014, the Ministry of Education and Training (MOET) adopted a six-level foreign language proficiency framework compatible with the levels in the Common European Framework of Reference for Languages (CEFR). Students graduating from high school are required to achieve the A2 level so that they "can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc.; deal with most situations likely to arise while traveling; produce simple connected texts on topics which are familiar or of personal interest, describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans" (Vietnam Ministry of Education and Training, 2014).

On surface, the NHSGE English test itself is believed to be not adequate to evaluate the students' proficiency as it comprises fifty multiple-choice questions and does not evaluate a student's speaking, listening or writing skills. Students and parents often point fingers saying that the existing curricula often followed in secondary and high schools seem to be outdated, lack a clear understanding of the fast changing globalized work environment, and under-prepare the students to meet the international standards. Further, due to the lack of investments, the rural schools mostly face the brunt of not having qualified English teachers.

Learning English, having a high proficiency and doing well in the NHSGE English test also highlight the glaring social divide between the 'haves' and 'havenots'. Affluent families, especially in large urban areas, such as HCMC, Hanoi and Hai Phong, send their children to expensive private after-school tutorial centers which often hire foreign tutors from native English speaking countries. On the other hand, pupils in remote rural areas rely mostly on the school time for English learning. Not surprisingly about two-third of these students score below 5, indicating insufficient basic proficiency. If this trend is allowed to continue without any immediate interventions from the national, provincial and local policy makers, then the urban-rural divide will continue to expand which may bring more unintended social maladies. Those living in cities, and generally proficient in English, will gravitate more toward those jobs and higher education where English plays a decisive role, thereby effectively excluding those from rural areas, and hence possibly widening the aforementioned social inequalities.

The objective of the study

Though the above two subsections provide plenty of anecdotal evidence that something is amiss in English teaching in Vietnamese schools, and there is a glaring disparity between the rural and urban students in terms of the NHSGE English test score, it is about the high time that we employ some rigorous mathematical / statistical approaches to study this aspect which touches the lives of hundreds of thousands of students. Our objective is to look at the students' collective performance at a more granular level, that is - the provincial level, and see what factors are contributing to this performance.

The existing research gap, the research purpose, and the research questions

(1) The purpose of this research study is to collect the existing data over a period of three years (2019 - immediately pre Covid-19 pandemic, 2020 - during the height of Covid-19 pandemic, and 2021- immediately post Covid-19 pandemic) and observe the patterns through aggregate provincial English test score data. Not only did we (the researchers) want to see the discernible trends in the data, but also figure out if these trends could be explained by some key factors. In the absence of any existing quantitative study (relevant to Vietnam), we tried a few measurable variables (or, factors) which are thought to hold some significance in explaining the aggregate provincial trends.

(2) The research questions centered around not only identifying the key factors (if any) affecting the aggregate provincial trends, but also quantifying their impacts, so that the policy makers could focus on those factors to enhance the overall students' performance.

2. LITERATURE REVIEW

As indicated earlier, to the best of our knowledge, there does not exist any rigorous statistical research to understand the Vietnamese high school students' English test performance. Any substantial educational policy change and/or initiative must follow a thorough understanding of the ground realities. Otherwise, it may result not only in wasting the precious resources, but also may cause more harm than any good.

3. MATERIALS AND METHODS

(1) Vietnamese National High School Examination dataset in 2019, 2020, 2021 (English score).

(2) The databases on the aforementioned variables were extracted from Vietnam Statistical Year Books (General Statistics Office of Vietnam, 2019; 2020; 2021).

To understand the high school students' English test score better, the list of 63 provinces of Vietnam (which includes 59 designated provinces and four major cities each of which enjoys the status of a province) is presented first. Tables A.1 and A.2 in the Appendix show the list of all the provinces along with the mean (i.e, average) and the median (i.e., the central-most value) English test scores for the three recent years (i.e., year 2021, 2020 and 2019).

In our quest to quantify the provinces' overall English test performances, we have studied several possible explanatory variables data on which were readily available. Ultimately, it has been found that two variables as described below have had the most significant contributions:

(1) Per Capita Gross Domestic Product - PCGDP: monthly average income per capita (unit: 10 million Vietnamese Dong (VND)); and

(2) Average Number of Students per High School (ANSHS).

Yearly, the PCGDP of a province is calculated by the province's official GDP figure divided by the provincial population size. On the other hand, a province's ANSHS is determined by a province's total number of high school students (taking the English test) divided by the total number of high schools located in that province.

4. RESULTS AND DISCUSSION

4.1. The Macro-level Big Picture: Descriptive Statistics

Remark 4.1. *Note that Table A.1 and A.2 go from 1 through 64. This is to be consistent with the earlier numbering of the provinces. Also, the item-20 on those lists is missing because the province (Ha Tay) has been merged with the city ("province") Hanoi thereby resulting into 63 current provinces.*

Figures A.1 - A.3 in the Appendix show the more descriptive box-plots of all the 63 provinces for the three years mentioned above. For each province, the corresponding horizontal line shows the range of most of the students belonging to that province and the horizontal box represents the middle half (the middle 50%) of the group ranging from the first quartile to the third quartile with the median value (shown with a cut mark) in between. The few dots above and below each box, toward the end points (or boundaries), indicate the potential outliers. It is evident that the remote rural provinces are at the lower spectrum while the predominantly urban ones are at the higher end.

Remark 4.2. *See the tell-tale sign of differences between what happened in 2019-2020 and in 2021. First of all, Figure 1 clearly shows that while there was a slight bump under the right tail (of 2019 and 2020) indicating the nationwide high performing students, it became more pronounced in 2021. This is because - even though the Covid-19 pandemic started affecting the rest of the world beginning January 2020, it barely touched Vietnam in the entire 2020 due to its strong initial public health measures. But Vietnam couldn't escape the second wave of Covid-19 Delta variant that ravaged the country starting in late March 2021 which witnessed the country going into a complete lockdown and bringing the economic activities, especially in the south, to a virtual standstill. Only in late September 2021 the country started lifting the lockdown slowly, in a phased manner, to ease the economic hardship. This second wave of Covid not only caused a large number of casualties for which the country was ill-prepared, but also impacted the academic activities, which went fully in remote mode of instructions, thereby negatively and disproportionately affecting the rural students. While urban and well-to-do families weathered this pandemic storm over six months (March to September, 2021) relatively better, the rural and/or financially struggling families were not so fortunate. Notice that the first four provinces (1-4) pretty much stayed ahead of all others as a whole during the three years of the study period. In 2019 and 2020, almost all provinces had a bunch of over-achievers (marked by the dots along the vertical line, indicating their positions as potential outliers within each respective province) which vanished in 2021 (except for a handful few). In the 2021 national exam, which took place right in the middle of the heavy pandemic lockdown, the first four provinces (Hanoi, Hai Phong, Da Nang and HCMC) - the four major (and perhaps privileged) urban centers with significant populations (with almost a fifth of the country's total population) pulled ahead of the rest of the country with their relatively better financial footing. Most likely this resulted in a clear double peak in the 2021 test score histogram (see Figure 1).*

Remark 4.3. *We also looked at other possible covariates such as, - the number of teachers per thousand students (essentially "teacher-student ratio"), number of teachers per high school, etc., to explain the provincial mean and/or median test score, but our subsequent data analyses didn't find them having any significant contribution, and hence not worth pursuing.*

The rest of the paper is organized as follows. In Section 4.2, we present our multiple linear regression models which establish a direct relationship between a province's mean English test score (PM-ETS) with its PCGDP and ANSHS. The same has been done for a province's median English test score (PMd-ETS). Our rigorous statistical model verification leads to an improved (and more technical) model which can provide a better future prediction, and this has been covered in Section 4.3. Finally, the paper ends with some comments and recommendations.

4.2. The Micro-level Granular Picture: Connecting the dots through Regression

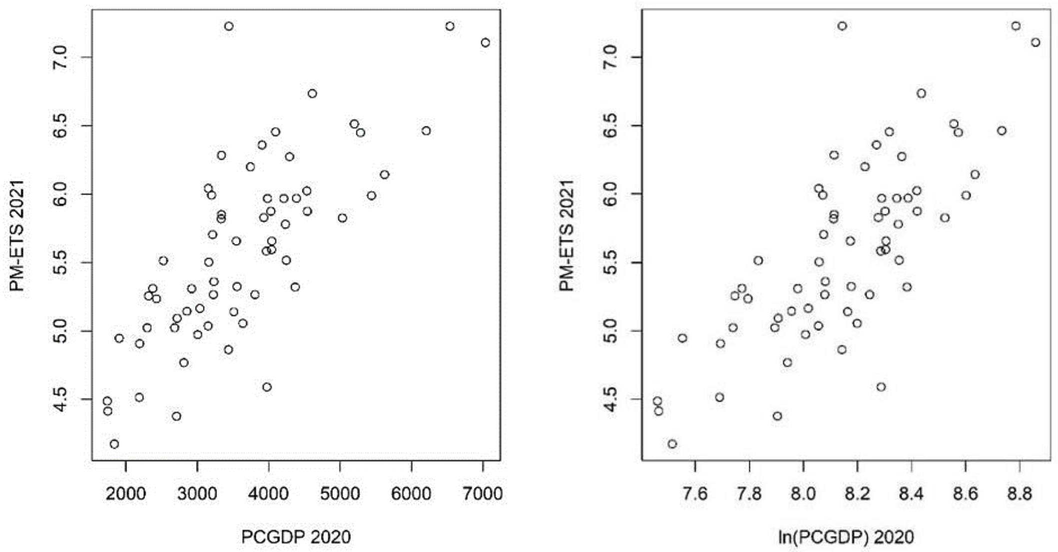


Figure 2. PM-ETS 2021 plotted against PCGDP 2020 as well as ln(PCGDP) 2020

As mentioned in the previous section, the mean (or average) English test score of a province (i.e., PM-EST) can be explained to some extent by the two explanatory variables, namely PCGDP and ANSHS (see the paragraph before Remark 4.3). (We have done a similar modeling for the provincial median English test score, i.e., PMd-ETS, but to keep our focus fixed we have explained the matter in detail for PM-ETS.) To see the effect of each of these two explanatory variables on PM-ETS ('Provincial Mean English Test Score'), the following figures are instrumental. Figures 2 shows the plots of PM-ETS against PCGDP as well as ln(PCGDP). When a variable is widely dispersed, a natural logarithm ('ln') transformation can bring it 'in line' by condensing its scale, and may offer a better linearity when another (dependent) variable is plotted against it.

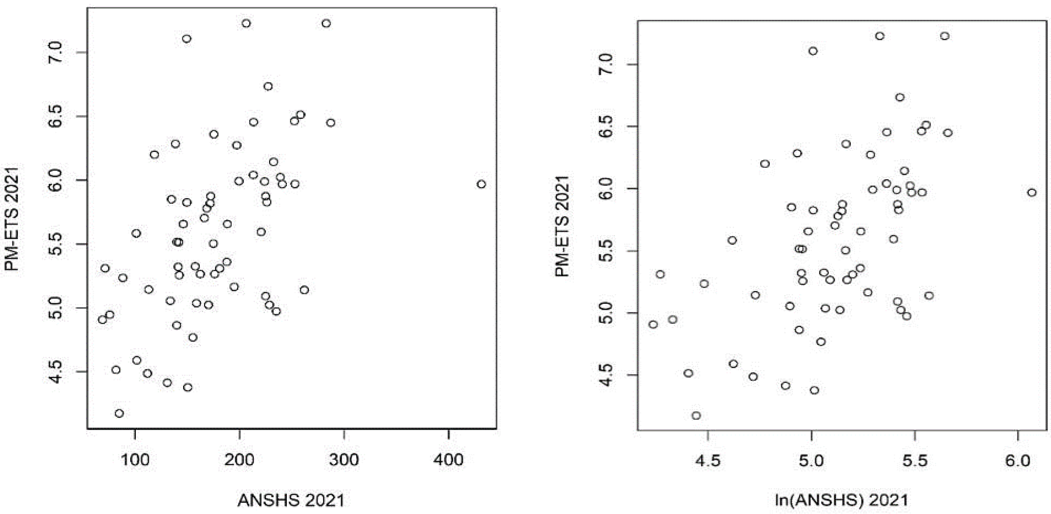


Figure 3. PM-ETS 2021 plotted against ANHS 2020 as well as ln(ANSHS) 2020

***Remark 4.4**. We believe that the student's overall performance in a province (as expressed through PM-ETS) in a particular year is being determined by the PCGDP of the previous year. That is why Figure 2 shows the plots of 'PM-ETS 2021' against 'PCGDP 2020' as well as 'ln(PCGDP) 2020'. On the other hand, Figures 3 shows the plots of 'PM-ETS 2021' against 'ANSHS 2021' and 'ln(ANSHS) 2021'. As such, ANSHS is not expected to change drastically from year to year, but yearly PCGDP can alter greatly as we have witnessed before and during the pandemic time.*

As Figures 2 - 3 show and our more detailed statistical analyses have revealed, there is not much gain in using the logarithmic transformation of the variables PCGDP and ANSHS. Therefore, for the sake of simplicity and easier comprehension, our regression models have used the original explanatory variables, i.e., based upon the plots in Figures 2 and 3.

Adopting the standard notations in a regression model set up, define (for a given year, say t)

$Y =$ PM-ETS (for a province in year t)

$X_1 =$ PCGDP (for a province in year $(t-1)$)

$X_2 =$ ANSHS (for a province in year t)

Then using the data from all the $n = 63$ provinces, we start with the full linear regression model as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \varepsilon \qquad (4.1)$$

where each 'β' represents the regression coefficient associated with the corresponding variable, and the term $(X_1 X_2)$ in (4.1) is the interaction between X_1 and X_2. The 'error' in the above model, i.e., ε, is assumed to follow $N(0, \sigma^2)$ distribution.

As a demonstration, we present the results for the year $t = 2021$ here. Using the data from $n = 63$ provinces, the above model (4.1) yields the estimated regression coefficients as follows: $\widehat{\beta}_0 = 3.766$, $\widehat{\beta}_1 = 4.143$, $\hat{\beta}_2 =$

1.949×10^{-3} and $\hat{\beta}_{12} = -3.823 \times 10^{-4}$. However, the resultant Analysis of Variance (ANOVA) table (see Table A.3 in the Appendix) shows that the coefficient β_{12} can be taken as zero (since $H_0: \beta_{12} = 0$ is retained against the alternative $H_A: \beta_{12} \neq 0$ with a p-value of 0.962). In other words, the interaction term $(X_1 X_2)$ has a negligible effect in explaining Y when both X_1 and X_2 are already present in the model. This yields the reduced model as follows:

$$\text{PM-ETS} = \beta_0 + \beta_1(PCGDP) + \beta_2(ANSHS) + \varepsilon \qquad (4.2)$$

where both the terms are found to be 'significant'. The detailed ANOVA table for the model (4.2) has been provided in Appendix A.3. This reduced model provides R-square ≈ 0.59, adjusted R-square ≈ 0.58, and the multiple correlation coefficient of approximately 0.768 (which is the Peasron's correlation coefficient between the left hand side (LHS) and the right hand side (RHS) of (4.2) with β - coefficients replaced by their estimates). The approximate R-square value of 0.59 implies that about 59% of the total variability in the observed PM-ETS values in 2021 across 63 provinces can be attributed due to the corresponding provincial values of PCGDP and ANSHS. The R-square value, in a way, quantifies the amount of (combined) impact of the explanatory variables (such as PCGDP and ANSHS) on the explained variable (that is, PM-ETS) through the multiple linear regression model (4.2).

An integral part of further inferences based on the regression model fitting is to check the model assumptions. In this regard, the observed residuals are used to verify the (a) normality of the model errors; and (b) homoscedasticity, i.e., equality of error variances across all provinces.

For the 2021 PM-ETS study, the normality assumption seems inconclusive as the two standard test methods (Anderson - Darling test (ADT) and Shapiro - Wilk test (SWT)) yield p-values of 0.20 and 0.03 respectively. While the ADT p-value is sufficiently large (compared to the usual threshold of 0.05), and supports the normality assumption for the error term in (4.2), the SWT's p-value looks "small", thereby challenging the validity of normality.

Not being able to satisfy the normality assumption conclusively as mentioned above is a partial setback for the regression model since all subsequent statistical inferences (such as interval estimation of the regression coefficients and/or undertaking subsequent hypothesis testing) depend heavily on this assumption. However, as far as the model fitting is concerned, the estimated parameters are obtained by the least squares method. Perhaps this setback is not surprising given the fact that the residual values do show a bit skewed form as the following Figure 4 shows.

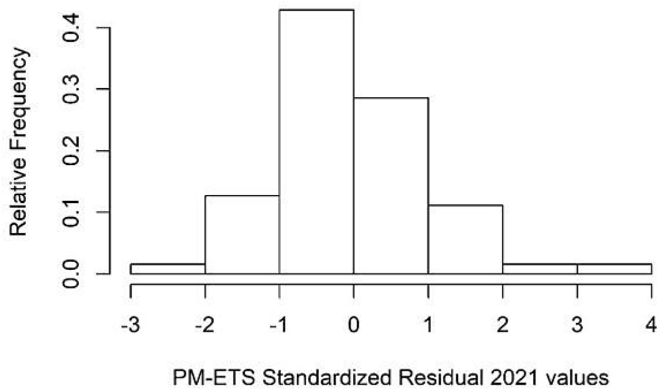


Figure 4. Relative frequency histogram of standardized residuals for the model (4.2)

Therefore, to improve over the model (4.2), we propose a different quantitative measure of a province's English test performance. This is done by using the 'Provincial Median English Test Score' or PMd-ETS instead of PM-ETS. When a variable is asymmetrically distributed, its median is a more stable measure of the center than the mean. This is because the mean is more prone to be affected by the extreme values than the median.

We repeat the model (4.1) where Y now plays the role of PMd-ETS with everything else remaining the same.

Again, the full model can be reduced further by dropping the insignificant interaction term. The final model, with a marginally better R-square and Adjusted R-square comes out as follows:

$$\text{PMd-ETS} = \beta_0 + \beta_1(PCGDP) + \beta_2(ANSHS) + \varepsilon \qquad (4.3)$$

where β - coefficients have been estimates as $\hat{\beta}_0 = 2.881$, $\widehat{\beta}_1 = 5.647, and\ \hat{\beta}_2 = 2.337 \times 10^{(-3)}$. Also, R-square= 0.59, Adjusted R-square= 0.58 and multiple correlation coefficient $\simeq 0.771$. (Note that there hasn't been any change in R-square or Adjusted R-square for the year 2021).

The model assumption tests for 2021 PMd-ETS yield the following results: the normality assumption turns out to be inconclusive again as the p-values for ADT and SWT come out to be 0.20 and 0.004, respectively. However, as for the equality of variances assumption, the standard test (Modified Levene's test (MLT)) produces a large p-value of 0.378, thereby retaining the homoscedasticity assumption.

The residual values for 2021 PMd-ETS do not seem to be normally distributed as the same pattern persists. The following Remark 4.5 and Table 1 summarizes our findings regarding PM-ETS and PMd-ETS.

Remark 4.5. *For the years 2021 and 2020, the use of PMd-ETS seems to work a little better than the corresponding PM-ETS as far as the regression model fitting is concerned. Yet this is not so for all the three years as their performances tend to reverse in 2019 which we will see later. Therefore, we present the fitted regression models for both PM-ETS as well as PMd-ETS with all the details in Table 1.*

Remark 4.6. *For the normality assumption, it has been noted that ADT and/or SWT are yielding extremely small p-values for all the three years. Both the tests do not provide sufficiently large p-values simultaneously, not even for a single year, which could have accepted normality conclusively. This implies that the normality assumption is not tenable. However, for the equality of variances assumption, the standard test (Modified Levene's test (MLT)) produces a large p-value, thereby retaining the assumption.*

Remark 4.7. *The overall micro-level trends revealed by the results presented in Table 1 is quite interesting. Note that the estimated coefficients of X_1 and X_2 are both positive. This means that a province can improve its mean (as well as median) English test score by improving its per capita income (i.e., enhancing the overall economic prosperity) as well as increasing the number of students per school. Both of these characteristics have already been seen in the top ranking four provinces (see Tables A.1 and A.2). Also, as the number of students goes up in a high school, contrary to the common fear of having stretched resources, it actually creates a more competitive academic environment which perhaps propels the students to excel more and thereby improving the overall aggregate English test score.*

Table 1. Summary of regression analysis for PM-ETS and PMd-ETS for three years

$Y = \beta_0 + \beta_1(PCGDP) + \beta_2(ANSHS) + error$		
Year	Y = PM-ETS	Y = PMd-ETS
2021	$\beta_0 = 3.791, \beta_1 = 4.075, \beta_2 = 1.086 \times 10^{-3}$	$\beta_0 = 2.881;\ \beta_1 = 5.647, \beta_2 = 2.337 \times 10^{-3}$
	$R_{sq} = 0.5904; Adj\ R_{sq} = 0.5767$	$R_{sq} = 0.5944; Adj\ R_{sq} = 0.5809$
	Mult.Corr.Coeff = 0.768	Mult.Corr.Coeff = 0.771
	ADT P-value = 0.2	ADT P-value = 0.2
	SWT P-value = 0.03	SWT P-value = 0.004
	MLT P-value = 0.54	MLT P-value = 0.38
2020	$\beta_0 = 2.721, \beta_1 = 3.094, \beta_2 = 3.008\text{x}10^{-3}$	$\beta_0 = 2.631;\ \beta_1 = 3.286, \beta_2 = 1.476\text{x}10^{-3}$
	$R_{sq} = 0.4969; Adj\ R_{sq} = 0.4801$	$R_{sq} = 0.5296; Adj\ R_{sq} = 0.5139$
	Mult.Corr.Coeff = 0.705	Mult.Corr.Coeff = 0.728
	ADT P-value ≈ 0	ADT P-value = 0.0707
	SWT P-value ≈ 0	SWT P-value = 0.0004
	MLT P-value = 0.81	MLT P-value = 0.8142
2019	$\beta_0 = 2.840, \beta_1 = 3.164, \beta_2 = 1.664 \times 10^{-3}$	$\beta_0 = 2.881;\ \beta_1 = 3, \beta_2 = 1.4 \times 10^{-3}$

	$R_{sq} = 0.5184; Adj\ R_{sq} = 0.5024$	$R_{sq} = 0.4417; Adj\ R_{sq} = 0.4231$
	Mult.Corr.Coeff $= 0.761$	Mult.Corr.Coeff $= 0.665$
	ADT P-value $= 0.166$	ADT P-value $= 0.069$
	SWT P-value $= 0.002$	SWT P-value $= 0.0028$
	MLT P-value $= 0.41$	MLT P-value $= 0.5634$

4.3. Further Improvement of the Regression Model

In the earlier subsection, we have seen that the regression model of the formula

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \qquad (4.4)$$

played a pivotal role in explaining the overall trend of ETS. Once again, note that X_1 and X_2 represent PCGDP and ANSHS respectively, and Y can be taken either as PM-ETS (which yields the specific model (4.2)) or PMd-ETS (which produces the specific model (4.3)) for the three study years (2019, 2020 and 2021). Table1 shows that the two explanatory variables X_1 and X_2 have been able to explain approximately half of the total variability in Y through the structure (4.2) or (4.3), which relies on the assumption that ε follows $N(0, \sigma^2)$ distribution. However, as seen from the relative frequency histograms of the observed residuals (which looked quite skewed), and corroborated by the formal normality tests (SWT and ADT), the normality assumption doesn't look convincing for the given data. As such, this violation doesn't pose any serious problem for estimating the regression coefficients (i.e., β_i's) by the ordinary least squares (OLS) method, but it can create problems in further inferences, especially for predicting the value of Y when X_1 and X_2 are known. The three years for which we collected the data looked quite diverse; - 2019 was a pre-pandemic year, 2020 was the beginning of the pandemic which affected the rest of the world but had a minimal impact on Vietnam, and 2021 was in a full-blown pandemic which ravaged Vietnam miserably. So, suppose we take a future year which may look similar to one of these three years. Then, for that future year, say t^*, let the variable Y be denoted by Y^*, and X_1 and X_2 be denoted by X_1^* and X_2^* respectively (This new notation system has been adopted for convenience to differentiate between the past and the future). Then Y^* is predicted by

$$\hat{Y}^* = \hat{\beta}_0 + \hat{\beta}_1 X_1^* + \hat{\beta}_2 X_2^* \qquad (4.5)$$

where $\hat{\beta}_i$ $(i = 0,1,2)$ are available from one of the past years which may look closest to the future year t^* that we are interested in. (Note that Y^* can play the role of either PM-ETS or PMd-ETS of the future year t^*.) The predicted value $\hat{Y}^*$ in (4.2) is obtained by using the fact that $E(Y^*) = \beta_0 + \beta_1 X_1^* + \beta_2 X_2^*$, since $E(\varepsilon^*) = 0$ due to the normality assumption. However, the performance of the predictor $\hat{Y}^*$ (in (4.5)) may not be appealing if ε doesn't follow the normality. Usually, the performance of a predictor $\hat{Y}^*$ is evaluated by its Prediction Mean Squared Error (PMSE), defined as $E(\hat{Y}^* - Y^*)^2$, or by its Prediction Mean Absolute Error (PMAE), defined as $E|\hat{Y}^* - Y^*|$.

Due to potential nonnormality of the error (ε^*) in the predictive model

$$Y^* = \beta_0 + \beta_1 X_1^* + \beta_2 X_2^* + \varepsilon^*, \qquad (4.6)$$

we can assume a skew normal distribution (SND) for ε^* with location parameter 0, scale parameter σ and skew parameter λ (henceforth called as "$\varepsilon^* \sim SND(0, \sigma, \lambda)$"). SND is a generalization of the usual normal model, and the skew parameter λ, which can take any real value, allows SND to take both positively skewed (if $\lambda > 0$) as well as negatively skewed (if $\lambda < 0$). Further, $SND(0, \sigma, \lambda)$ becomes $N(0, \sigma^2)$ if $\lambda = 0$. There is a rich literature on SND, and any interested reader can see Thiuthad and Pal (2019) and the references therein. The flexibility that SND provides over the normal distribution comes at a heavy cost. While most of the widely used sampling distributions related to a normal distribution are fairly known analytically, it is not so for an SND, and this poses a serious challenge when it comes to subsequent inferences. Here, under the SND assumption, one has to rely mainly on the computational results.

Though the SND error model as prescribed in (4.6) gives a good flexibility over the normal error model, estimation of the model parameters (i.e., β_i's, σ and λ) is a challenging task. Note that, using the SND error model, all the model parameters need to be estimated for the past data (i.e., for each of the years of 2019, 2020 and 2021) before we attempt to use the predictive model. Unlike the OLS method universally used (which also happens to be the maximum likelihood estimation (MLE) method) under the normality assumption, there are three estimation techniques under the SND model, namely - $OLS + MME$ (a combination of the OLS and the method of moments estimation (MME) method), maximum penalized likelihood estimation ($MPLE$), and the MLE (using a special centralized reparametrization). Since these are highly technical tools, and requires substantial theoretical background, details are going to be omitted here. Any interested reader can see the details provided in two recent papers - Huynh et al. (2021a and 2021b).

The following Tables 2 and 3 provide the estimated parameters under the SND error model for regressing PM-ETS as well as PMd-ETS yearwise.

Table 2. Estimated parameters under SND error for PM-ETS

Year	Estimated Parameters	Estimation Method		
		OLSE+MME	MPLE	MLE
	$\hat{\beta}_0$	2.0595	2.2574	2.1715
	$\hat{\beta}_1$	0.0006	0.0006	0.0006
2021	$\hat{\beta}_2$	0.0023	0.0024	0.0026
	$\hat{\sigma}$	1.0311	0.8397	0.8992
	$\hat{\lambda}$	5	2.1828	2.9591
	$\hat{\beta}_0$	2.0583	2.3246	2.3125
	$\hat{\beta}_1$	0.0003	0.0003	0.0003
2020	$\hat{\beta}_2$	0.0015	0.0012	0.0010
	$\hat{\sigma}$	0.5004	0.5377	0.5747
	$\hat{\lambda}$	5	1.4512	1.8558
	$\hat{\beta}_0$	2.0080	2.2163	2.1860
	$\hat{\beta}_1$	0.0003	0.0003	0.0003
2019	$\hat{\beta}_2$	0.0014	0.0012	0.0012
	$\hat{\sigma}$	0.0014	0.0012	0.0012
	$\hat{\lambda}$	5	2.1550	2.7179

Table 3. Estimated parameters under SND for PMd-ETS

Year	Estimated Parameters	Estimation Method		
		OLS+MME	MPLE	MLE
	$\hat{\beta}_0$	3.2467	3.3372	3.2593
	$\hat{\beta}_1$	0.0004	0.0004	3.2592
2021	$\hat{\beta}_2$	0.0018	0.0020	0.0021
	$\hat{\sigma}$	0.4875	0.5975	0.6426
	$\hat{\lambda}$	4.4975	1.8086	2.4187
	$\hat{\beta}_0$	3.3594	3.1282	3.1668
	$\hat{\beta}_1$	0.0003	0.0003	0.0003
2020	$\hat{\beta}_2$	0.0030	0.0026	0.0026
	$\hat{\sigma}$	0.6213	0.5755	0.5982
	$\hat{\lambda}$	-5.0000	-1.2290	-1.4113
	$\hat{\beta}_0$	2.3019	2.4397	2.3774
	$\hat{\beta}_1$	0.0003	0.0003	0.0003
2019	$\hat{\beta}_2$	0.0017	0.0016	0.0017
	$\hat{\sigma}$	0.4425	0.5472	0.5881
	$\hat{\lambda}$	5.0000	2.0860	2.8520

Remark 4.8. *The main difference between the normal error model and the SND error model is the presence of the skew parameter λ in the latter. (When $\lambda = 0$, the SND boils down to normal distribution.) Hence, the estimated value of λ is extremely important in the performance of the predictive model (4.3) where $\varepsilon^* \sim SND(0, \sigma, \lambda)$. However, as shown in Table 2, the three estimation techniques produce quite diverse sets of estimated model parameters. While all but λ estimated values are somewhat similar, the estimated λ looks quite different under the three estimation methods, and this has a profound implication for the prediction of Y^* when X_1^* and X_2^* are given. Unlike the normal error model which yields only one predictor as in (4.2), the SND error model yields three predictors as follows one of which happens same as in (4.2):*

$$\hat{Y}^{*(1)} = \hat{\beta}_0 + \hat{\beta}_1 X_1^* + \hat{\beta}_2 X_2^* + \hat{\gamma} \qquad (4.7)$$

$$\hat{Y}^{*(2)} = \hat{\beta}_0 + \hat{\beta}_1 X_1^* + \hat{\beta}_2 X_2^* + \hat{\sigma} m_0(\hat{\lambda}) \qquad (4.8)$$

$$\hat{Y}^{*(3)} = \hat{\beta}_0 + \hat{\beta}_1 X_1^* + \hat{\beta}_2 X_2^* + \hat{\sigma} M_0(\hat{\lambda}) \qquad (4.9)$$

Where $\gamma = \sqrt{\frac{2}{\pi}}\frac{\sigma\lambda}{\sqrt{1+\lambda^2}}$, $m_0(\lambda) = \eta_\lambda - (\gamma_1/2)\sqrt{1-\eta_\lambda^2} - (sign(\lambda)/2)exp(-2\pi/|\lambda|)$, with $\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}, \eta_\lambda = \sqrt{2/\pi}\delta, \gamma_1 = (2-\pi/2).(\sigma\sqrt{2/\pi})^3(1-2\delta^2/\pi)^{-3/2}$ and $M_0(\lambda)$ is the solution of the following equation for y with $T(y \mid \lambda) = \{\Phi(y) - 0.5\} / 2$, where $T(y|\lambda)$ is the Owen's T-function in λ given as

$$T(y \mid \lambda) = (1/2\pi)\int_0^\lambda \left[\exp\left\{-y^2(1+x^2)/2\right\}\right]/(1+x^2)dx$$

As stated earlier, $\hat{Y}^{*(1)}$ happens to be same as $\hat{Y}^*$ (in (4.2)). Highly technical details of (4.7) - (4.9) can found in Huynh et al., 2021(b).

With the three estimation methods (OLS+MME, MPLE and MLE), thus we now have a total of nine predictors of Y^*, out of which $\hat{Y}^{*(1)}$ is always same as $\hat{Y}^*$ in (4.5) irrespective of the estimation method. Therefore, effectively, the SND error model produces six new competitors to $\hat{Y}^*$ for prediction purposes. How good are these new predictors can only be studied by investigating the PMSE and PMAE through a detailed parametric bootstrap method (see Huynh et al., 2021b, Sec.3). Out of the three estimation methods, the OLS+MME technique shows the overall best performance (for the given data) as shown in the following Figures 10 - 13 where PMSE and PMAE have been plotted against λ.

Remark 4.9. *What do the Figures 5 - 8 tell us? The plots of PMSE and PMAE tell us how the SND model can help us in improved prediction of the value of Y when the values of X_1 and X_2 are known (or anticipated). The standard regression model assumes that the model error follows a normal distribution (as presented in sub-section 4.2). But if we assume a more flexible SND in lieu of a normal model for the error term, then the SND skew parameter λ is an extra parameter which needs to be estimated from the data. For example, take the case where Y represents PM-ETS. The unknown λ has been estimated by three estimation techniques which yield the estimated values as 2.1828 (by MPLE), 2.9591 (by MLE) and 5.000 (by OLS+MME). While the actual λ is unknown, the aforementioned estimates tell us that the true value of λ ought to be somewhere (roughly) from 2 to 5. Now let us look at the predictors in (4.8) and (4.9) which are available only under the SND. The predictor in (4.7), which is one of the three predictors under SND is also the sole predictor which we obtain under the normal model. Figure 5 shows the PMSE curves of the three predictors, and note that as λ deviates from 0, the predictors (4.8) and (4.9) are having lower PMSE than that of the predictor (4.7) at λ = 2, 3, 4 and 5, the predictor (4.8) provides approximately (25.90)%, (36.42)%, (41.84)% and (45.00)%, respectively, improvements over the predictor (4.7).*

On the other hand, these improvement values for the predictor (4.9) are *(9.41)%*, *(12.90)%*, *(14.66)%* and *(15.66)%*, respectively. Similarly, Figure 6 shows the similar trends when Y represents PMd-ETS. Figures 7 and 8 are likewise when the criterion PMAE is used instead of PMSE.

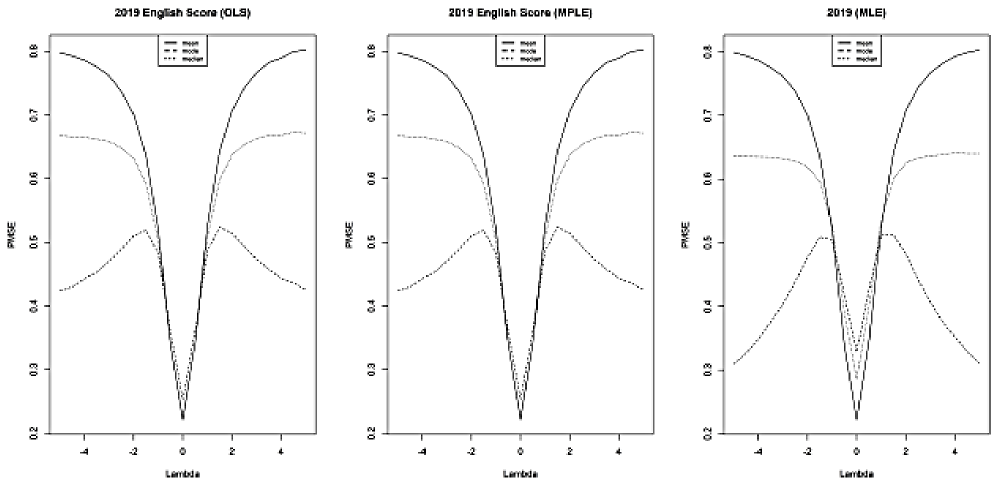


Figure 5. The plots of PMSE of the three predictors for three different methods (OLS+MME, MPLE and MLE) for PM-ETS in 2019

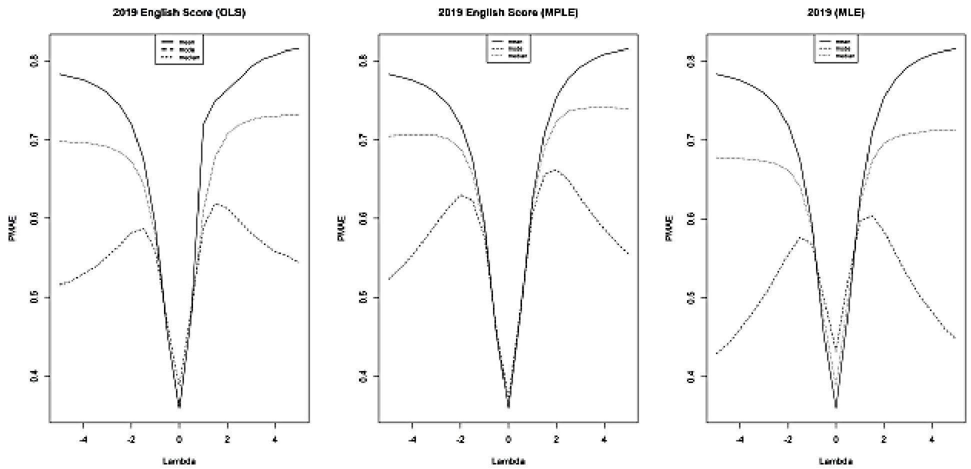


Figure 6. The plots of PMAE of the three predictors for three different methods (OLS+MME, MPLE and MLE) for PM-ETS in 2019

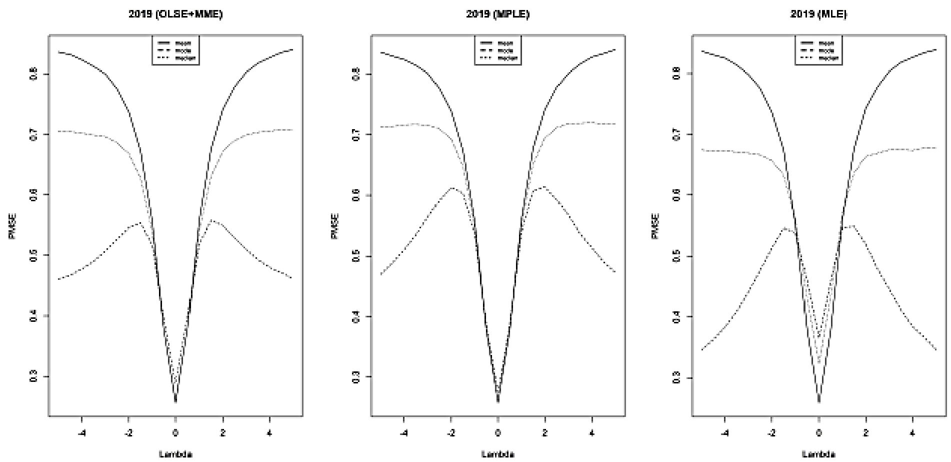


Figure 7. The plots of PMSE of the three predictors for three different methods (OLS+MME, MPLE and MLE) for PMd-ETS in 2019

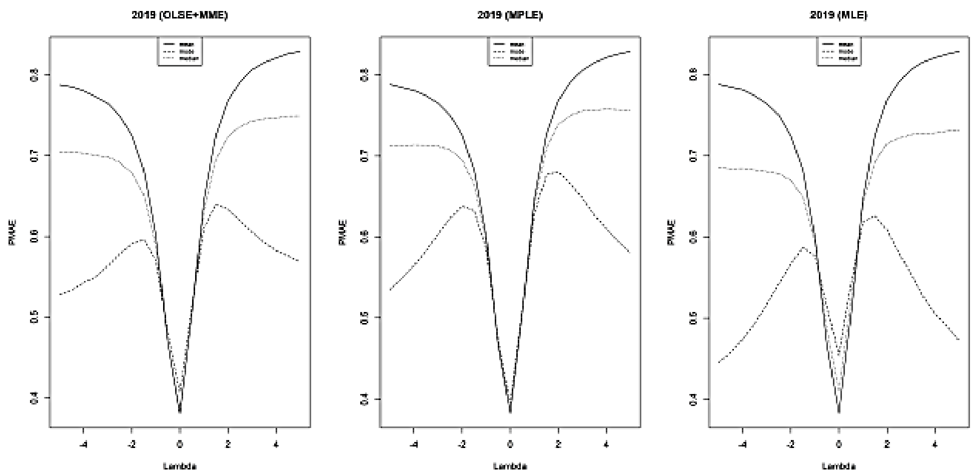


Figure 8. The plots of PMAE of the three predictors for three different methods (OLS+MME, MPLE and MLE) for PMd-ETS in 2019

5. CONCLUSION

In this study we have tried to explain the aggregate English test performance of a province (as measured by the provincial mean English test score (PM-ETS) as well as the provincial median English test score (PMd-ETS). Interestingly, two key factors have been found to account for nearly half of the total variability (from province to province) of that aggregate performance. This has been demonstrated by the analysis of the English test score (ETS) data over a period of three recent consecutive years using a suitable multiple linear regression model. The two factors which appear to play important roles are essentially measuring the overall economic prosperity of a province (as measured by PCGDP), and the overall academic competitiveness faced by the students (as measured by ANSHS). Both of these two factors (or covariates) are positively associated with the provincial aggregate English test performance. In other words, as PCGDP and/or ANSHS for a province go(es) up (or down), the aggregate provincial performance also goes up (or down).

Purely on a technical side, our study has been divided in two components: one in the form of a multiple linear regression model under the usual normal error assumption (covered in sub-section 4.2), and another in a more advanced multiple linear regression under the SND error model (as shown in Subsection 4.3, which is a generalization of subsection 4.2). Both the aforementioned subsections indicate that the estimated coefficients of PCGDP as well as ANSHS are strictly positive, thereby indicating that the overall students' aggregate performance can be enhanced by improving the province's (that they live in) economic well-being as well as improving the overall academic competitiveness. This latter finding may have some other deeper implications, and ANSHS may be confounded with some other factors which are yet to be identified. However, essentially it says that just reducing the number of students per high school may not be helpful as far as aggregate performance is concerned. Probably the quality of the teachers and other resources (like library / digital facilities) do play an important role in enhancing the students' overall performance (but we do not have any data yet to back up this assumption).

Future extension of this work is possibly in two potential directions as follows. (a) Focusing on the students' performance (i.e., ETS) at a more granular level, that means, perhaps one may look at the district or ward level, and at the same time at the corresponding local level PCGDP and ANSHS. But while the latter covariate may be observable at a local level, similar information on local level PCGDP may be hard to access. (b) Bringing in other useful covariates into the model which can help us understand the students' aggregate performance in a more meaningful way. As mentioned earlier, the existing multiple linear regression model using provincial PCGDP as well as provincial ANSHS (as shown in subsection 4.2) can explain about half (or a little more than half) of the total variability in provincial aggregate ETS. We must find other measurable covariates, which could be qualitative as well as quantitative, that can enhance our model's explicability of the aggregate ETS. Educators, policy makers and sociologists may propose such useful explanatory variables for which provincial and/or local level data ought to be available that can strengthen the statistical model.

We sincerely hope that this study will generate further interest among the educators as well as policy makers not only in Vietnam but also in other Southeast Asian nations with similar socio-economic and cultural backgrounds. It is also hoped that the ensuing debates and discussions can help with identifying other explanatory variables (apart from PCGDP and ANSHS) which can improve the regression model further in order to study the aggregate ETS performance of the high school students. Last but not least, this statistical analysis has been based upon the available "observational study" data. Is it possible to improve the students' aggregate ETS performance through some "experimental study"? For example, by altering the existing English learning curricula by requiring the students to watch English news channels and/or asking them to read English print media on a regular basis which can possibly help the students achieve a higher level of English proficiency? This is a more serious research question all should be concerned about.

One final word of caution: The statistical analysis undertaken in this study has been done purely as an academic research study by the authors to understand the prevailing condition of the Vietnamese high school students' performance in terms of English test score (ETS). Neither does it reflect the views of the academic institutions that the authors belong to, nor should it be construed as an indictment against any existing educational system.

6. APPENDIX (We note that the province listed as 20 has been merged with Hanoi)

*Table A1. List of the 63 provinces along with the **mean** English test scores for the years 2021, 2020 & 2019*

Code	Provinces	mean English test scores			Code	Provinces	mean English test scores		
		2021	2020	2019			2021	2020	2019
1	Ha Noi	6.46	5.18	5.03	33	Quang Nam	5.32	3.99	3.95
2	HCMC	7.23	5.87	5.79	34	Quang Ngai	5.50	1.63	4.09
3	Hai Phong	6.51	5.02	4.70	35	Kon Tum	5.31	4.22	4.06
4	Da Nang	6.45	5.10	4.94	36	Binh Dinh	7.23	5.87	5.79
5	Ha Giang	4.17	3.24	3.08	37	Gia Lai	5.26	4.12	3.96
6	Cao Bang	4.51	3.66	3.38	38	Phu Yen	5.26	4.21	4.00
7	Lai Chau	4.95	3.90	4.00	39	Dak Lak	5.09	3.73	3.94
8	Lao Cai	5.24	3.99	3.74	40	Khanh Hoa	6.04	4.73	4.62
9	Tuyen Quang	5.02	3.73	3.33	41	Lam Dong	6.20	4.85	4.66
10	Lang Son	5.02	3.87	3.33	42	Binh Phuoc	5.59	4.47	4.31
11	Bac Kan	4.91	3.67	3.70	43	Binh Duong	7.11	5.51	5.18
12	Thai Nguyen	5.27	4.04	3.81	44	Ninh Thuan	5.31	4.09	3.95
13	Yen Bai	5.51	4.17	3.68	45	Tay Ninh	5.52	4.44	4.32
14	Son La	4.41	3.36	2.99	46	Binh Thuan	5.97	4.74	4.69
15	Phu Tho	5.82	4.53	4.49	47	Dong Nai	6.14	4.86	4.68
16	Vinh Phuc	6.27	4.83	4.55	48	Long An	5.78	4.59	4.34
17	Quang Ninh	5.87	4.31	3.90	49	Dong Thap	5.58	4.43	4.23
18	Bac Giang	5.83	4.49	4.07	50	An Giang	6.29	4.91	4.71
19	Bac Ninh	5.99	4.57	4.33	51	Ba Ria - Vung Tau	6.74	5.28	5.11
20	Hai Duong	5.97	4.51	4.24	52	Tien Giang	6.02	4.81	4.65
21	Hung Yen	5.66	4.27	3.96	53	Kien Giang	5.32	4.13	3.92
22	Hoa Binh	4.38	3.40	3.29	54	Can Tho	5.83	4.62	4.47
23	Ha Nam	5.87	4.55	4.35	55	Ben Tre	5.66	4.42	4.35
24	Nam Dinh	6.46	5.12	4.89	56	Vinh Long	5.99	4.64	4.42
25	Thai Binh	5.97	4.62	4.34	57	Tra Vinh	4.86	4.04	3.86
26	Ninh Binh	6.36	4.87	4.49	58	Soc Trang	5.05	4.01	3.82
27	Thanh Hoa	5.14	3.88	3.68	59	Bac Lieu	5.85	4.63	4.33
28	Nghe An	4.97	3.99	3.75	60	Ca Mau	5.17	4.12	3.91
29	Ha Tinh	5.36	4.08	3.74	61	Dien Bien	4.49	3.67	3.68
30	Quang Binh	5.04	3.87	3.67	62	Dak Nong	4.77	3.78	3.75
31	Quang Tri	5.14	4.12	4.03	63	Hau Giang	4.59	3.72	3.59
32	Thua Thien-Hue	5.70	4.46	4.27					

*Table A2. List of the 63 provinces along with the **median** English test scores for the years 2021, 2020 & 2019*

Code	Provinces	median English test scores			Code	Provinces	median English test scores		
		2021	2020	2019			021	020	2019
1	Ha Noi	6.8	4.8	4.4	33	Quang Nam	5.0	3.6	3.6
2	HCMC	7.8	5.8	5.6	34	Quang Ngai	5.2	4.0	3.8
3	Hai Phong	6.8	4.8	4.2	35	Kon Tum	4.8	3.8	3.6
4	Da Nang	6.8	4.7	4.6	36	Binh Dinh	7.8	5.8	5.6
5	Ha Giang	3.6	3.0	2.8	37	Gia Lai	4.8	3.6	3.6
6	Cao Bang	3.8	3.2	3.2	38	Phu Yen	4.8	3.8	3.6
7	Lai Chau	4.8	3.8	4.0	39	Dak Lak	4.6	3.4	3.6
8	Lao Cai	4.8	3.6	3.4	40	Khanh Hoa	6.0	4.4	4.2
9	Tuyen Quang	4.6	3.4	3.0	41	Lam Dong	6.2	4.6	4.4
10	Lang Son	4.6	3.6	3.0	42	Binh Phuoc	5.2	4.2	4.0
11	Bac Kan	4.4	3.4	3.4	43	Binh Duong	7.4	5.4	5.0
12	Thai Nguyen	4.8	3.6	3.4	44	Ninh Thuan	5	3.8	3.6
13	Yen Bai	5.2	3.8	3.4	45	Tay Ninh	5.2	4.2	4.0
14	Son La	4.0	3.2	2.8	46	Binh Thuan	5.8	4.6	4.4
15	Phu Tho	5.4	4.2	4.2	47	Dong Nai	6.2	4.6	4.4
16	Vinh Phuc	6.2	4.4	4.2	48	Long An	5.6	4.4	4.0
17	Quang Ninh	5.6	3.8	3.4	49	Dong Thap	5.4	4.2	4.0
18	Bac Giang	5.6	4.2	3.8	50	An Giang	6.2	4.8	4.6
19	Bac Ninh	5.8	4.2	4.0	51	Ba Ria - Vung Tau	7.0	5.0	4.8
20	Hai Duong	5.8	4.2	3.8	52	Tien Giang	6.0	4.6	4.4
21	Hung Yen	5.4	3.8	3.4	53	Kien Giang	5.0	4.0	3.6
22	Hoa Binh	3.8	3.0	3.0	54	Can Tho	5.6	4.4	4.2
23	Ha Nam	5.6	4.2	4.0	55	Ben Tre	5.4	4.2	4.2
24	Nam Dinh	6.4	4.8	4.6	56	Vinh Long	5.8	4.4	4.2
25	Thai Binh	5.8	4.2	4.0	57	Tra Vinh	4.4	3.8	3.6
26	Ninh Binh	6.2	4.6	4.0	58	Soc Trang	4.6	3.8	3.6
27	Thanh Hoa	4.6	3.4	3.2	59	Bac Lieu	5.6	4.4	4.0
28	Nghe An	4.4	3.6	3.4	60	Ca Mau	4.8	4.0	3.6
29	Ha Tinh	5.0	3.8	3.4	61	Dien Bien	4.2	3.4	3.4
30	Quang Binh	4.6	3.4	3.4	62	Dak Nong	4.2	3.4	3.4
31	Quang Tri	4.6	3.6	3.6	63	Hau Giang	4.2	3.4	3.4
32	Thua Thien-Hue	5.4	4.0	3.8					

Table A3. PM-ETS 2021 ANOVA table

	df	Sum Sq	Mean Sq	F value	P-value
PCGDP	1	16.815 (57.01%)	16.815	83.511	≈ 0
ANSHS	1	0.597 (0.02 %)	0.597	2.966	0.902
Residuals	60	12.081 (40.96 %)	0.201		

Table A4. PM-ETS 2020 ANOVA table

	df	Sum Sq	Mean Sq	F value	P-value
PCGDP	1	11.940 (44.47 %)	11.940	53.041	≈ 0
ANSHS	1	1.400 (5.22 %)	1.400	6.217	0.015
Residuals	60	13.507 (50.31 %)	0.225		

Table A5. PM-ETS 2019 ANOVA table

	df	Sum Sq	Mean Sq	F value	P-value
PCGDP	1	9.988 (49.86 %)	9.988	62.111	≈ 0
ANSHS	1	0.398 (1.98 %)	0.398	2.476	0.121
Residuals	60	9.648 (48.16 %)	0.161		

Table A6. PMd-ETS 2021 ANOVA table

	df	Sum Sq	Mean Sq	F value	P-value
PCGDP	1	31.808 (57.63 %)	31.808	85.2406	≈ 0
ANSHS	1	1.000 (1.81 %)	1.000	2.6807	0.1068
Residuals	60	22.389 (40.56 %)	0.373		

Table A7. PMd-ETS 2020 ANOVA table

	df	Sum Sq	Mean Sq	F value	P-value
PCGDP	1	11.876 (51.49 %)	11.876	65.676	≈ 0
ANSHS	1	0.337 (1.46 %)	0.337	1.864	0.1773
Residuals	60	10.850 (47.05 %)	1.808		

Table A8. PMd-ETS 2019 ANOVA table

	df	Sum Sq	Mean Sq	F value	P-value
PCGDP	1	8.988 (42.85%)	8.9880	46047	≈ 0
ANSHS	1	0.277 (1.32%)	0.2768	1.418	0.2384
Residuals	60	11.712 (55.85%)	0.1952		

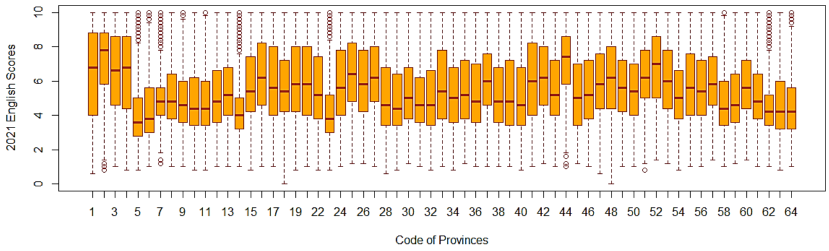


Figure A1. Box-plot of the provincial English test scores for the year 2021

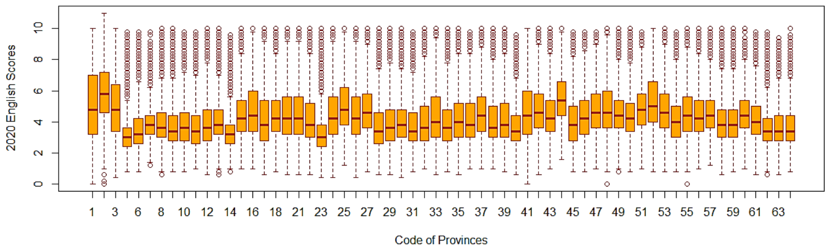


Figure A2. Box-plot of the provincial English test scores for the year 2020

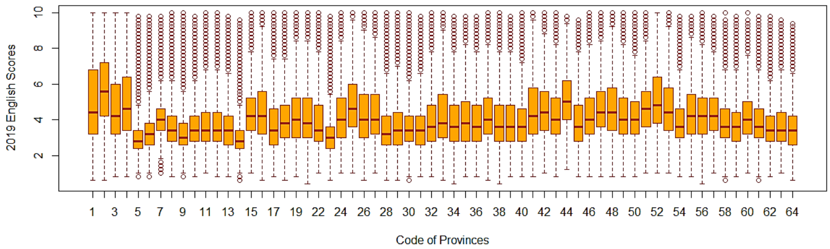


Figure A3. Box-plot of the provincial English test scores for the year 2019

Conflict of Interest: No potential conflict of interest relevant to this article was reported.

Acknowledgements: The second author's research has been funded by the University of Economics and Law, Vietnam National University Ho Chi Minh City/VNU-HCM.

REFERENCES

General Statistics Office of Vietnam (2019). *Statistical Year Book*. Vietnam Statistics Publishing House.

General Statistics Office of Vietnam (2020). *Statistical Year Book*. Vietnam Statistics Publishing House.

General Statistics Office of Vietnam (2021). *Statistical Year Book*. Vietnam Statistics Publishing House.

Huynh, U., Pal, N., Truong, C., & Nguyen, M. (2021a). A Statistical Profile of Arsenic Prevalence in the Mekong Delta Region. *Thailand Statistician*, *19*(2), 339-360.

Huynh, U., Pal, N., Truong, C., & Nguyen, M. (2021b). Regression Model under Skew-Normal Errors with Applications in Predicting Groundwater Arsenic Level in the Mekong Delta Region. *Environmental and Ecological Statistics*, *28*, 323-353. https://doi.org/10.1007/s10651-021-00488-2

Thuithad, P., & Pal, N. (2019). Point estimation of the location parameter of a skew-normal distribution: Some fixed sample and asymptotic results. *Journal of Statistical Theory and Practice*, *13*(2), 13-37. https://doi.org/10.1007/s42519-018-0033-4

Vietnam Ministry of Education and Training (2014). *Circular No. 01/2014/TT-BGDDT dated January 24, 2014 on promulgating the six-level Foreign Language Proficiency Framework for Vietnam*.