Mining Educational Data to Analyze the Student's Performance in TOEFL iBT Reading, Listening and Writing Scores

Khaled M. Hassan[1]
Demonstrator at ISA Information System, International Smart Association, International Smart Association, ISA, Nasr City, Egypt

Mohammed Helmy Khafagy[2]
Professor of Computer Science Computer Science Department Faculty of Computers and Information, Fayoum University Fayoum

Mostafa Thabet[3]
Lecturer of Information System, Information System Department Faculty of Computers and Information, Fayoum University Fayoum

***Abstract*—Student scores in TOEFL IBT reading, listening, and writing may reveal weaknesses and deficiencies in educational institutions. Traditional approaches and evaluations are unable to disclose the significant information hidden inside the student's TOEFL score. As a result, data mining approaches are widely used in a wide range of fields, particularly education, where it is recognized as Educational Data Mining (EDM). Educational data mining is a prototype for handling research issues in student data which can be used to investigate previously undetected relationships in a huge database of students. This study used the EDM to define the numerous factors that influence students' achievement and to create observations using advanced algorithms. The present study explored the relationship among university students' previous academic experience, gender, student place and their current course attendance within a sample of 473 (225 male and 248 female). Educational specialists must find out the causes of student dropout in TOEFL scores. The results of the study showed that the model could be suitable for investigation of important aspects of student outcomes, the present research was supposed to use the statistical package for social sciences (SPSS V26) for both descriptive and inferential statistics and multiple linear regressions to improve their scores.**

Keywords*—Educational data mining; students score; linear regression; TOEFL; Statistics

I. Introduction

Over the last decade, test developers and experts have fixated much of their time and focus on developing a theoretical view of language ability in order to understand better the nature of language proficiency, as well as developing and applying more sophisticated statistical tools to analyze language tests and test takers' performance in order to best tap these issues[1]. However, language testing research shows that language aptitude is not the only factor influencing test takers' performance. Almost all screening processes in academic environments, from seeking college admission to applying for an exchange student programmer, require the applicant to present TOEFL iBT or other Standard English language test scores.

The TOEFL iBT (Test of English as a Foreign Language) Language testing is largely concerned with whether the results clearly effectively reflect test takers' underlying ability in a certain area in a given testing setting [2]. After graduation, English proficiency is necessary for developing career options and attaining aspirational goals in the workplace [3]. The Educational Testing Service (ETS) commissioned a recent survey study and found a high link between high English proficiency and the income of young professionals (full-time workers in their 20s or 30s) across all major industries. This higher income allows them to put more money into improving their English abilities, which are "a vital instrument for success in today's world". Test-takers personality factors to the testing scenario, such as education level, Gender, and place, can all affect their performance [4]. But these construct-irrelevant elements are regarded as potential causes of test bias, which might cause the acquired results to be unrepresentative of the underlying skill that a language test is attempting to assess. As a result, a thorough assessment of the likely effects of such factors is worthwhile.

Taking these factors into account and the popularity of the TOEFL iBT as a proficiency exam worldwide, this study aims to determine the future effects of test education level, Gender, and place on TOEFL iBT listening reading and writing results.

II. Literature Survey

Test fairness is a challenging topic in the literature when it comes to language testing. Debates about test fairness aim to create tests free of discrimination and contribute to testing equity [5, 6]. When students with the same language ability perform differently on a test, it may be called discriminatory. When the substance of the test is discriminatory to test takers from certain groups, other criteria such as education level, Gender, and test place play a factor. The test's requirements may have different impacts on test takers from different groups; test taker factors such as education level, Gender and place can all contribute to test bias.

These factors can impact a test's validity and lead to measurement mistakes. As a consequence, in the design and

development of language exams decreasing the impact of these factors that are not part of the language competence is a top objective [7].

The association between TOEFL score and GPA was shown to be positive and statistically significant; however, it was less for engineering students than for students in other professions and for engineering courses than for non-engineering courses. In logistic regressions of CAE pass rate and graduation rate, the TOEFL score was also statistically significant, showing an increased probability of success with a higher TOEFL score. However, model goodness-of-fit values were low, showing that many students defied overall trends in their performance [8].

Accord to the previous survey, a mixed ANOVA was used to answer the following study question: Is there a significant difference between pre and post TOEFL test scores for male and female students? Is there an interaction between male and female students' pre and post TOEFL test scores? According to those findings, there was a substantial change between pre and post TOEFL exam scores, but no significant variation between genders. Furthermore, no correlation was found between male and female students' pre and post TOEFL test scores [9].

In agreement with the past research, there was a relationship between overseas students' academic performance and their language skills, academic self-concept and other factors that influence academic achievement. The research looked at first-year international students enrolled in undergraduate business programs at a Canadian English-medium institution. The following data was gathered on the students: grades in degree program courses, annual GPA, and EPT scores (including sub scores).

Students also filled out an academic self-concept measure. In addition, instructors in two obligatory first-year business courses were interviewed regarding the academic and linguistic requirements in their courses and the profile of successful students to acquire additional information about success in first-year business courses [10].

In the other side the purpose of this study was to determine whether there was a significant difference in the capacity of male and female students to respond to factual and vocabulary-in-context questions on the TOEFL-like reading comprehension test. The results of reading comprehension tests taken from twenty-one male and twenty-one female students in the English Education Program were used for secondary data analysis. Through the use of random sampling, samples were chosen. Utilizing an independent sample t-test, data were evaluated [11].

On the other hand in this study, the self-efficacy of university students in responding to TOEFL questions is examined in relation to gender and participation in TOEFL courses. This study uses a descriptive design with a total sample of 200 university students from two large institutions who are majoring in both English and non-English [12].

III. Proposed Methodology

After reviewing data and determining the research aim and objectives, this paper examines the effects of characteristics such as education level, attendance, and student gender to examine students' scores in TOEFL iBT reading, listening, and writing using data mining approaches. For this study's techniques and data preparation procedures, methodologies are discussed below.

A. Dataset

The data for this study came from 473 students. Arabic is one of their first languages. 473 students in total took the TOEFL. The study enlisted the participation of 225 male and 248 female students (Table I).

TABLE I. Attributes of the Dataset

Attributes	**Details**
Gender	Male, Female
Education level	Faculty
Place	Cairo, Sheikh Zayed
Attendance	Number of Course Attendance

B. Data Preparation

All activities were taken from the raw data to create the final dataset (data that was entered into the design tool). The dataset's variables were prepared to generate the models needed in the next phase.

The students received a variety of English language skills, including a TOEFL preparation session, during the rigorous English language program. The TOEFL scores of the students were used as the research tool. At the end of the course, students take the TOEFL (paper-based test). Students were in class for five hours a day and were given TOEFL-related assignments. Listening, grammar/structure, and reading are the three skills that make up the TOEFL score. The TOEFL score ranges between 310 and 677. This study aims to determine the future effects of test education level, gender, place, and attendees on TOEFL iBT listening, reading, and writing results.

IV. Model and Algorithm

Fig. 1 depicts a framework for predicting student success. First, the data on student performance is fed into this system. This student data set has been preprocessed to eliminate noise and make the data set more consistent. The input data set is then subjected to various SPSS statistics analyses. Next, data analysis is carried out. Finally, different algorithms' categorization results are compared.

Likewise, gender is another factor that is usually studied, but there is a lack of good research to identify whether male and female language learners have significantly different TOEFL results. From a psychological standpoint, there are numerous variables related to gender [13].

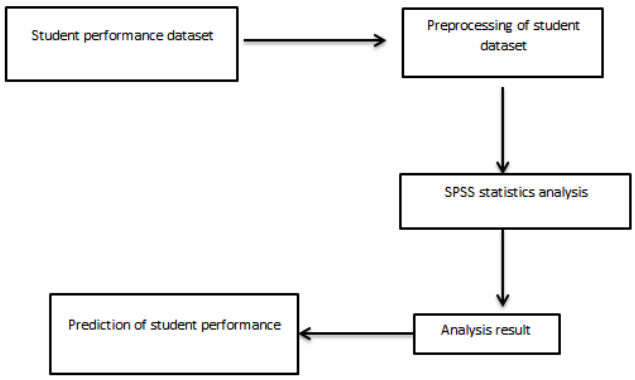

Fig. 1. A Framework for Student Performance Prediction.

In general, females are believed to be more successful in language learning than males. Therefore, many scholars in language acquisition studied how gender disparities can affect students' language learning proficiency. In other words, ten studies found that female students were superior to male students in reading comprehension. In contrast, five studies found that male students were superior [14,15] also undertook a quantitative study to see if there are any gender differences in TOEFL scores and found no significant differences. The Educational Testing Service (ETS), on the other hand, came to a different result.

According to the survey, female pupils are more advanced than male students [16]. Females, for example, outperformed males in writing and reading, though the difference was minor. On the other hand, Male students performed higher in terms of listening and comprehension, as well as vocabulary proficiency [17].

Additionally, a standardized English language assessment examination, such as the Test of English as a Foreign Language, is required at most English language colleges and universities (TOEFL). However, because there are few standardized evaluation measures for all candidates, English proficiency ratings are occasionally utilized for purposes other than evaluating the "abilities of non-native English speakers to use and understand English."

However, in the lack of standard ranking techniques for all candidates, the TOEFL score may be used as a stand-in for those criteria; the TOEFL score is occasionally employed as a predictor of how well a potential student will perform at a university. Even when the TOEFL is not used as the main measure of academic success, minimum TOEFL score requirements are frequently enforced.

Despite the fact that the underlying English-language communication abilities that TOEFL scores represent may be significantly more important to academic performance in specific areas, TOEFL score minimums for admission frequently do not vary among academic majors or fields of study. Requiring the same minimum TOEFL score whatever of a student's selected major may lead to the exclusion of otherwise talented students from academic programmers where academic achievement is not contingent on language competence [18]. For example, an increased TOEFL score is less correlated with academic success in college students than in other college students (possibly because English communication skills largely determine academic success in these areas). It may be reasonable to adopt the TOEFL score entry requirements. More lenient for engineering applicants, especially those who can show enough preparation through means other than a TOEFL score.

Despite the fact that course enrollment has tripled in the past 10 years, little is known about the impact of environment tests and attendance on learning. According to a recent study of college students, course attendance and the student place have an impact on the examination scores. Therefore, differences in student accomplishment between groups should be viewed with caution. This study adds to the body of knowledge by addressing a recurring problem of earlier research: determining the impact of various classroom test conditions on exam scores. The features of test environments are rarely described in previous research. This study compares test scores from students who took examinations off-campus with test scores from students who were called back to school for probationary exams within a semester [8].

V. Experiments and Results

The analysis of this paper was done using the statistical package for social sciences (SPSS V26) for both descriptive and inferential statistics. In this work, ANOVA was used as a statistical analysis method. Because this study examines the significance of group differences, it uses an ANOVA statistical model with a continuous dependent variable (TOEFL scores) and categorical independent factors.

Because this study tries to observe the interaction between gender differences, ANOVA is the most appropriate statistical procedure among the numerous varieties of ANOVA [19]. Pre and post TOEFL scores are within-subject factors, while male and female are between-subject variables. To address the first study question, a statistically significant mean difference between before and post TOEFL scores will be studied. After that, we'll look at the statistically significant mean difference between male and female TOEFL scores. The impacts will next be compared between the TOEFL scores of males and females.

Table II provides descriptive statistics for the selected variables, including the minimum (Min), maximum (Max), mean (M), standard deviation (SD), and coefficient of variation (CV) (M=48.36,SD=7.519,CV=1 5.55%),(M=47.24,SD= 7.972,CV=16.88%),(M=47.07,SD=8.354,CV=17.75%), (M=475.38,SD=70.869,CV=14.91 %) respectively.

Table II shows some descriptive statistics and bivariate correlations among the selected variables provided in this section.

TABLE II. Descriptive Statistics

	N	Min	Max	Mean	SD
Listening	473	24	68	48.36	7.519
Grammar	473	27	68	47.24	7.972
Reading	473	27	67	47.07	8.354
Total	473	300	653	475.38	70.869

TABLE III. MULTIPLE CORRELATIONS

		Listening	Grammar	Reading	Total
Listening	***Pearson Correlation***	1			
	P-value				
	N	473			
Grammar	***Pearson Correlation***	.647***	1		
	P-value	.000			
	N	473	473		
Reading	***Pearson Correlation***	.642***	.781***	1	
	P-value	.000	.000		
	N	473	473	473	
Total	***Pearson Correlation***	.847***	.911***	.909***	1
	P-value	.000	.000	.000	
	N	473	473	473	473

Table III displays the bivariate correlations between the study's primary variables; all of the correlations were statistically significant at 0.001. These correlations vary between.642 and.642, indicating that all variables in the study have substantial moderate to strong multiple correlations.

Furthermore, the results of the multiple regression were reported, and it can be noticed that all variables have significant positive effect on the total score since (P<0.001), as a result, the null hypothesis is rejected, and the alternative hypothesis is accepted in Table IV.

TABLE IV. REGRESSION COEFFICIENT

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	***Std. Error***	***Beta***		
1	***(Constant)***	.334	1.234		.270	.787
	Listening	3.336	.032	.354	102.996	.000***
	Grammar	3.396	.038	.382	90.461	.000***
	Reading	3.257	.036	.384	91.499	.000***

*** P < 0.001

Table IV, the assumptions of this study were examined using multiple regression analysis in this part.

Table V, the F-test in ANOVA table confirms the significance of the model since (F=546827.6, P<0.001).

TABLE V. ANOVA TABLE

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	***Regression***	2363572.748	3	787857.583	52762.172	.000[b]
	Residual	7003.222	469	14.932		
	Total	2370575.970	472			

On the other side, the impact of demographic variables on the students' overall scores will be studied in this section. Finally, the normal distribution test was done utilizing Skewness and kurtosis tests to choose between parametric and nonparametric testing Table VI [20].

TABLE VI. TEST OF NORMALITY

	N	Skewness		Kurtosis	
		Statistic	***Std. Error***	***Statistic***	***Std. Error***
Total	473	.066	.112	-.756	.224

Table VI, the values of Skewness and kurtosis for the score were within the range of ±2, indicating that the total score was normally distributed, according to the normality statistics.

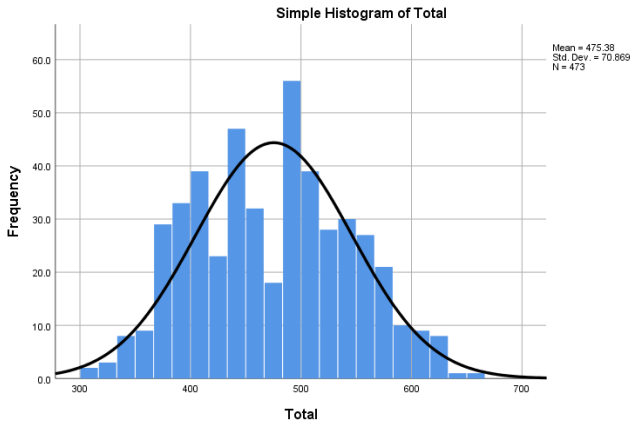

Fig. 2. Histogram and the Normal Curve of the Total Score.

Fig. 2 displys a normal distribution of data.

First hypothesis: there is a significant difference in total scores regarding the Gender of the students. The independent-samples t-test is the appropriate parametric test because Gender is a categorical variable with two independent categories.

Table VII, some descriptive statistics of the total score according to each category were given.

Fig. 3 can be concluded from that the average degree of females (487.49) was greater than that of males (462.04).

In addition, Levene's test for equality of variances was done and found that the variances were equal since ($F = .449, P > 0.05$). The results of the independent-sample t-test show that there is a significant difference in total scores between males and females since P-value is less than 0.05 as ($t = -3.961, P < 0.001$) Table VIII.

TABLE VII. DESCRIPTIVE STATISTICS OF THE TOTAL SCORES REGARDING THE GENDER

Gender	N	Min	Max	Mean	Std. Deviation
female	248	300	653	487.49	70.863
male	225	313	623	462.04	68.590
Total	473	300	653	475.38	70.869

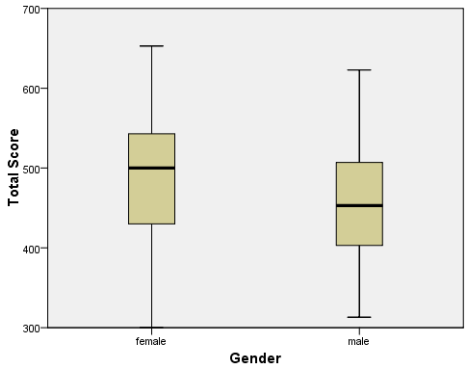

Fig. 3. Boxplot for the Total Scores of Students Regarding the Gender.

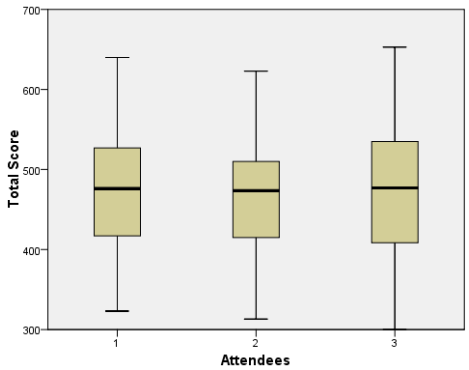

Fig. 4. Boxplot for the Total Scores of Students Regarding the Attendees.

TABLE VIII. INDEPENDENT SAMPLES T-TEST

Levene's Test for Equality of Variances			t-test for Equality of Means					
	F	*P-value*	*t*	*df*	*Mean*	*Std. Error*	*95% CI of the Difference*	
							Lower	*Upper*
Equal variances assumed	.449	.503	-3.961	471	-25.452	6.426	-38.078	-12.826
Equal variances not assumed			-3.967	469	-25.452	6.415	-38.058	-12.845

***P < 0.001

In Table VIII, the results of the independent-sample t-test show that there is a significant difference in total scores between males and females since P-value is less than 0.05 as$(t = -3.961, P < 0.001)$.

Moreover, in the second hypothesis: there is a significant difference in total scores regarding the attendees of the students. Since the student's attendance is a categorical variable with more than two independent categories, the suitable parametric test is the analysis of variance (ANOVA) test.

In Table IX, some descriptive statistics of the total score according to each category were given.

Fig. 4 presented that the average scores of students attending for the first time (477.19) was greater than that of the second time (474.39), and the third time (473.07).

In Table X, the results of the ANOVA test show that there is no significant difference in total scores between the number of attendees since the P-value is greater than 0.05 as$(F = .151, P > 0.05)$.

TABLE IX. DESCRIPTIVE STATISTICS OF THE TOTAL SCORES REGARDING THE ATTENDANCE

Attendees	N	Minimum	Maximum	Mean	Std. Deviation
1	226	323	640	477.19	71.650
2	124	313	623	474.39	65.787
3	123	300	653	473.07	74.747
Total	473	300	653	475.38	70.869

TABLE X. ANOVA TABLE

	Sum of Squares	Df	Mean Square	F	P-value
Between Groups	1525.638	2	762.819	.151	.860
Within Groups	2369050.333	470	5040.533		
Total	2370575.970	472			

As well the third hypothesis: there is a significant difference in total score regarding the place of the test. Since the place of the test is categorical variable with two independent categories, so the suitable parametric test is the independent-samples t-test.

Table XI shows some descriptive statistics of the total score according to each category were given.

Fig. 5 displayed that the average scores of students attend in Sheikh Zayed (484.81) were greater than those attending Cairo (466.38).

TABLE XI. DESCRIPTIVE STATISTICS OF THE TOTAL SCORES REGARDING PLACE OF THE TEST

place	N	Min	Max	Mean	Std. Deviation
Cairo	242	300	640	466.38	71.684
Sheikh Zayed	231	350	653	484.81	68.905
Total	473	300	653	475.38	70.869

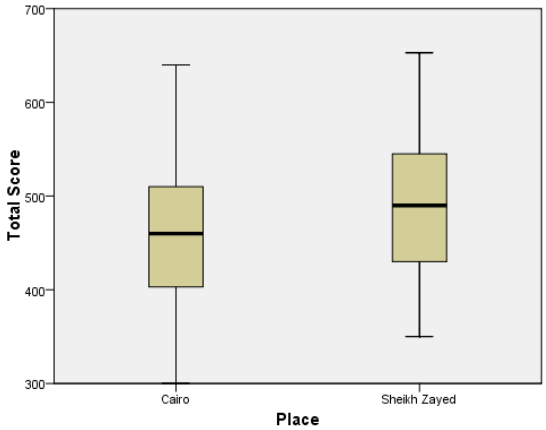

Fig. 5. Boxplot for the Total Scores of Students Regarding the Place of the Test.

TABLE XII. INDEPENDENT SAMPLES T-TEST

t-test for Equality of Means						
	t	df	Mean	Std. Error	95% CI of the Difference	
					Lower	*Upper*
Equal variances assumed	-2.848	471	-18.430	6.470	-31.144	-5.715
Equal variances not assumed	-2.851	471	-18.430	6.464	-31.132	-5.727

In Table XII, Levene's test for equality of variances reveals that the variances were equal since($F = .475, P > 0.05$). The results of the independent-sample t-test show that there is a significant difference in total scores between Cairo and Sheikh Zayed since P-value is less than 0.05 as ($t = -2.848, P < 0.01$).

Subsequently, the fourth hypothesis shows a significant difference in total scores regarding the level of education**.** Since the level of education is a categorical variable with more than two independent categories, the suitable parametric test is the analysis of variance (ANOVA) test.

TABLE XIII. DESCRIPTIVE STATISTICS OF THE TOTAL SCORES REGARDING THE LEVEL OF EDUCATION

Faculty	N	Min	Max	Mean	SD
Academy of Arts	3	300	570	465.67	145.074
African Institute	15	313	653	446.73	87.928
Agriculture	10	380	510	452.70	47.579
Applied Arts	6	417	577	499.00	50.931
Arab Academy	4	430	517	475.00	40.406
Archaeology	5	350	600	476.60	89.960
Arts	17	393	553	483.47	50.222
Commerce	40	363	600	473.05	59.843
Computer and Information	10	383	580	501.30	58.317
Dar Al Uloom	6	400	620	464.33	85.141
Dentistry	17	450	610	542.35	43.566
Economics and Political Sciences	13	360	617	523.38	72.240
Education	5	383	607	465.40	84.145
Egyptian fellowship	2	500	513	506.50	9.192
Engineering	25	423	640	519.80	53.275
environment institute	1	417	417	417.00	.
Georgia	13	413	623	528.00	49.427
Grant	2	450	560	505.00	77.782
industrial education	1	410	410	410.00	.
Institute of Arabic Studies	1	450	450	450.00	.
Institute of Technical healthy	109	350	523	411.31	39.451
Kindergarten	2	420	497	458.50	54.447
laser institute	1	413	413	413.00	.
Law	10	347	547	438.10	75.253
MBA	5	500	560	514.00	26.077
media	30	390	620	496.63	69.042
Medicine	27	410	623	549.19	45.668
National Institute of Intellectual Property	1	503	503	503.00	.
natural medicine	3	390	413	403.33	11.930
Naval Academy	7	377	507	477.71	49.291
Nursing	6	367	503	431.17	45.512
Oncology Institute	1	573	573	573.00	.
Pharmacy	14	447	627	542.21	52.850
Physical Education	5	327	473	406.60	55.383
Postgraduate Education	12	410	553	488.00	42.988
Research Institute	11	327	573	453.91	61.119
Sadat Academy	4	450	500	472.50	26.300
Sciences	16	450	597	520.81	39.507
Social Service	1	453	453	453.00	.
Statistics Institute	6	387	563	517.33	67.666
Tourism and Hotels	2	480	500	490.00	14.142
urban planning	2	557	587	572.00	21.213
veterinary medicine	2	460	610	535.00	106.066
Total	473	300	653	475.38	70.869

Table XIII shows some descriptive statistics of the total score according to each category were given.

Fig. 6 concluded that students' average scores were different across the level of education.

Table XIV shows the results of the ANOVA test show that there is a significant difference in total scores across the level of education since the P-value is less than 0.05 as ($F = 8.407, P < 0.001$).

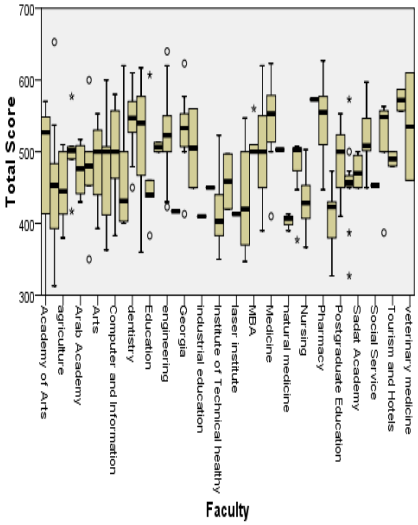

Fig. 6. Boxplot for Students' Total Scores Regarding the Level of Education.

TABLE XIV. ANOVA TABLE

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1068908.842	42	25450.211	8.407	.000
Within Groups	1301667.129	430	3027.133		
Total	2370575.970	472			

Finally, the fifth hypothesis: there is a significant difference in TOEFL parts (Listening, Grammar, and Reading) regarding the Gender.

Table XV shows some descriptive statistics of the TOFEL parts according to each category were given.

TABLE XV. DESCRIPTIVE STATISTICS OF THE TOEFL PARTS REGARDING THE GENDER

Gender		Listening	Grammar	Reading
female	***N***	248	248	248
	Minimum	24	27	28
	Maximum	68	68	67
	Mean	49.36	48.58	48.41
	Std. Deviation	8.060	7.869	7.932
male	***N***	225	225	225
	Minimum	32	27	27
	Maximum	68	67	67
	Mean	47.26	45.76	45.59
	Std. Deviation	6.720	7.838	8.572
Total	***N***	473	473	473
	Minimum	24	27	27
	Maximum	68	68	67
	Mean	48.36	47.24	47.07
	Std. Deviation	7.519	7.972	8.354

Fig. 7, can be concluded that for Listening, the average degree of females (49.36) was greater than that of males (47.26), for Grammar, the average degree of females (48.58) was greater than that of males (45.76), and for Reading the average degree of females (48.41) was greater than that of males (45.59).

Then, Levene's test for equality of variances was conducted. It can be noticed that for listening, we have unequal variances since $(F = 7.566, P < 0.01)$ but for Grammar, we have equal variances since$(F = .007, P > 0.05)$and the same for Grammar. We have equal variances since$(F = 1.870, P > 0.05)$. The results of the independent-sample t-test show that there is a significant difference in listening scores between males and females since P-value is less than 0.05 as$(t = -3.082, P < 0.01)$. Moreover, there is a significant difference in grammar scores between males and females since P-value is less than 0.05 as$(t = -3.900, P < 0.001)$. Finally, there is a significant difference in reading scores between males and females since P-value is less than 0.05 as$(t = -3.716, P < 0.001)$ Tables XVI and XVII.

In Tables XVI and XVII, since the Gender of the students is categorical variable with two independent categories; the suitable parametric test is the independent-samples t-test.

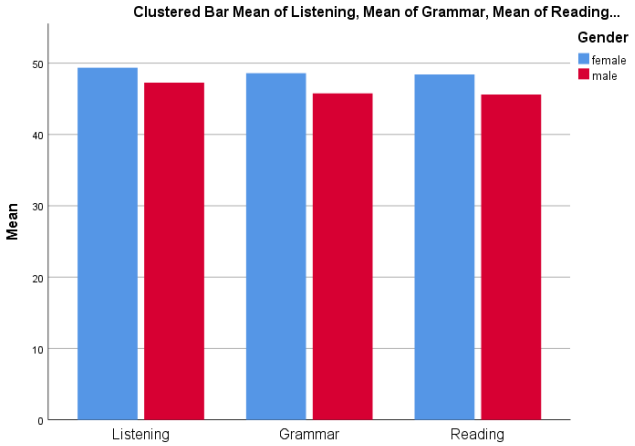

Fig. 7. Clustered Bar Chart for the TOEFL Parts Scores Regarding the Gender.

TABLE XVI. INDEPENDENT SAMPLES T-TEST

	Levene's Test for Equality of Variances		
		F	**P-value**
Listening	***Equal variances assumed***	7.566	.006
	Equal variances not assumed		
Grammar	***Equal variances assumed***	.007	.932
	Equal variances not assumed		
Reading	***Equal variances assumed***	1.870	.172
	Equal variances not assumed		

**P < 0.01

TABLE XVII. INDEPENDENT SAMPLES T-TEST

	t-test for Equality of Means							
		t	***df***	***P-value***	***Mean***	***Std. Error***	***95% CI of the Difference***	
							Lower	***Upper***
Listening	***Equal variances assumed***	-3.056	471	.002	-2.097	.686	-3.445	-.748
	Equal variances not assumed	-3.082	468	.002	-2.097	.680	-3.433	-.760
Grammar	***Equal variances assumed***	-3.900	471	.000	-2.820	.723	-4.241	-1.399
	Equal variances not assumed	-3.901	467	.000	-2.820	.723	-4.241	-1.399
Reading	***Equal variances assumed***	-3.716	471	.000	-2.820	.759	-4.311	-1.329
	Equal variances not assumed	-3.702	457	.000	-2.820	.762	-4.317	-1.323

**P < 0.01

VI. CONCLUSION

This study looked at the TOEFL results of 473 students based on how much time they spend studying, their educational level, gender, course attendance, and place. AS EXPECTED, the TOEFL scores improved from pre- to post-test, and the change was statistically significant. In this survey, there was significant difference by educational level, gender, attendance, and place difference. Furthermore, there was a relationship between male and female students' before and post TOEFL scores. As a result, the study's findings offer students with useful information. Furthermore, TOEFL educators can propose that the more time a student devotes to learning, the higher their TOEFL score will be. This also aids programmer makers in class design by giving them a sense of what students (who are prepared for the TOEFL) could expect. Because many students are applying to universities each year, generalizing TOEFL scores to the general population is insufficient.

REFERENCES

[1] BACHMAN, Lyle F., et al. Fundamental considerations in language testing. Oxford university press, 1990.

[2] WEIR, Cyril J. Language testing, and validation. Hampshire: Palgrave McMillan, 2005.

[3] Choi, I.-C, “The impact of EFL testing on EFL education in Korea. Language Testing”, Inn-Chull Choi, vol.25, No.1, pp.39–62, January 2008.

[4] Messick, S , “Validity and washback in language testing. Language Testing”, Samuel Messick, vol.13, No.3, pp. 241-256, November 1996

[5] Kunnan, A. J, “Test fairness, test bias, and DIF. Language Assessment Quarterly”, vol.4, No.2, pp. 109–112, Dec 2007.

[6] Llosa, Lorena, and Margaret E. Malone. "Student and instructor perceptions of writing tasks and performance on TOEFL iBT versus university writing courses." Assessing Writing, vol.34, pp. 88-99, October 2017.

[7] BACHMAN, Lyle F., et al. Language testing in practice: Designing and developing useful language tests. Oxford University Press, 1996.

[8] IW Wait, JW Gressel, “Relationship between TOEFL score and academic success for international engineering students”, Journal of Engineering Education, vol.98, No.4, pp. 389-398, October 2009.

[9] Saeun, L. E. E. "Improvement of pre-and post-tests and gender differences on TOEFL scores." Bulletin of Miyazaki Municipal University Faculty of Humanities, vol.25, No.1, pp. 193-204, 2018.

[10] Neumann, Heike, Nina Padden, and Kim McDonough. "Beyond English language proficiency scores: Understanding the academic performance of international undergraduate students during the first year of study." Higher Education Research & Development, vol.38, No.2, pp.324-338, Sep 2019.

[11] Destiyanti, Cahya, Muhammad Amin, and Lalu Jaswadi Putera. "Gender-Based Analysis of Students’ Ability in Answering Factual and Vocabulary-in-Context Questions of the TOEFL-Like Reading Comprehension Test." PALAPA vol.9, No.1, pp.1-17, 2021.

[12] Yoestara, Marisa, and Zaiyana Putri. "Gender and language course participation differences in the university students’ self-efficacy of TOEFL." Journal of Physics: Conference Series. Vol. 123,No.1, October2019.

[13] MACCOBY, Eleanor E.; JACKLIN, Carol Nagy. The psychology of sex differences. Stanford University Press, 1978.

[14] Hyde, Janet S., and Marcia C. Linn. "Gender differences in verbal ability: a meta-analysis." Psychological bulletin, vol.104, No.1, pp.53-69, Jul 1988.

[15] Lin, J., & Wu, F. Differential Performance by Gender in Foreign Language Testing, 2004.

[16] Cole, N. S. The ETS gender study: how females and males perform in educational setting. Princeton, NJ: Educational Testing Service, 1997.

[17] Boyle, J. P. “Sex Differences in Listening Vocabulary. Language Learning”,vol.37,No.2,pp.273-284,June 1987.

[18] Simner, M.L.” Use of the TOEFL as a standard for university admission: A position statement by the Canadian Psychological Association”. European Journal of Psychological Assessment, vol.14, No.3, pp. 261–65, 1998.

[19] Lomax, R. G. & Hahs-Vaughn . An introduction to statistical concepts (3rd Ed). New York, NY: Routledge, 2012.

[20] Gravetter, F., & Wallnau, L. Essentials of statistics for the behavioral sciences (8t h ed.). Belmont, CA: Wadsworth, 2014.