International Journal of Language Education
Volume 5, Number 3, 2021, pp. 207-223
ISSN: 2548-8457 (Print) 2548-8465 (Online)
Doi: https://doi.org/10.26858/ijole.v5i3.16200

A Shortcut in Language Testing: Predicting the Score for Paper-Based TOEFL Based on One Sub-Score

Samsul Anwar
Universitas Syiah Kuala, Indonesia
Email: samsul.anwar@unsyiah.ac.id

Faisal Mustafa
Universitas Syiah Kuala, Indonesia
Email: faisal.mustafa@unsyiah.ac.id

Received: 25 November 2020
Reviewed: 1 April 2021-1 June 2021
Accepted: 1 August 2021

Abstract

Using standardized tests such as paper-based TOEFL with three subtests for classroom assessment is restricted by the length of the test, which is usually longer than the class duration. Therefore, it is significant to be able to predict other subtests by conducting only one subtest. Therefore, the current study aimed to calculate prediction coefficients, enabling teachers to predict scores in paper-based TOEFL by conducting only one subtest. The data to create the prediction models were obtained from 2,030 scores of Institutional TOEFL, i.e. paper-based TOEFL without writing subtest. The prediction coefficient was calculated by using linear regression analysis. The result shows that the listening comprehension sub-score predicts the TOEFL score more accurately (MSE of 520) than other sub-scores (MSE of 553 and 587). The intercept for listening comprehension sub-score was 373.07, 357.14 for structure & written expression, and 364.19 for reading comprehension. In addition, the slope for each sub-score was 4.07, 5.96, and 4.63, respectively. Therefore, a listening test should be used in predicting the overall TOEFL scores for an accurate prediction.

Keywords: paper-based TOEFL; score prediction; linear regression model; language testing; correlational study

Introduction

A language test is an essential component in English language learning (Douglas, 2010). It is most commonly used to measure the students' language proficiency for placement, learning achievement, and diagnostic (J. D. Brown, 1996; Henning, 1987). Therefore, language test is usually included in the classroom syllabus. For a wider context, language test is a part of the curriculum for high schools in most countries where English is taught as a foreign or second language. The reliability of teachers made tests have been found to be lower compared to a standardized test (Coniam, 2009). EFL teachers who are concern about the reliability of a test or who prefer not to design their own test because it is a “time-consuming process” (Quaigrain & Arhin, 2017) can use standardized test material. One of the factors considered in selecting a language test is the length of the test (Bachman & Palmer, 1996). For teaching assessment in the classroom, the length of the test is restricted to the class duration. For standardized tests, the test ranges from 120 minutes (e.g. paper-based TOEFL), 165 minutes (e.g. IELTS), 200 minutes (e.g. TOEIC) to 240 minutes (e.g. internet-based TOEFL). The length of the test for standardized tests is an inhibiting factor for a classroom assessment. The class duration in high schools in most Asian countries is usually less than two hours (Shin & Kim, 2017; Williams, 2017), while the time required to complete a standardized test is at least two hours for the test and 30 minutes for the instruction (paper-based TOEFL without writing subtest). In addition, some tests cannot be conducted in some areas, such as listening test, due to noise, which is unpreventable in most school environment, unavailability of audio players or headphones. Thus, it is essential to predict the students' score for other subskills when only a score for a certain subskill is known. However, there has been no research which was aimed at making this prediction. Therefore, the current research was to calculate a prediction model for the standardized paper-based TOEFL when the score of only one subskill is known. Using the prediction model, a teacher can predict TOEFL scores when, for example, only structure score is known. The results of the study will particularly benefit English language teachers who prefer to use a paper-based TOEFL for classroom assessment.

Literature review

This section presents and discusses topics related to variables in this research, i.e. language testing, paper-based TOEFL, and prediction procedure in testing.

The significance of testing in language learning

In terms of its function, there are two types of tests or assessments involved in an educational context, i.e. formative and summative tests, both of which are significant for learning. A formative test is used to evaluate students during the learning process, while a summative test is conducted after a learning process to measure what students have learned (H. D. Brown, 2004). A famous example of formative assessment is a portfolio (Sulistyo et al., 2020). However, the current study is focused on the summative test, in which a standardized test is more applicable than in a formative assessment. According to Bailey (2017), a summative test can be used for instructional and diagnostic purposes. For both purposes, summative assessment tracks where a teacher should start, focusing on the areas which need more support (Green, 2014). Thus, it can be used as significant information in planning the instruction (Douglas, 2010). Therefore, in the context of language teaching, the test needs to be valid and reliable, and it has to accommodate all necessary language skills. In order that a test can satisfy such requirements, the test needs to be planned and piloted (Fulcher & Davidson, 2007). However, in the current practice, many summative tests are teachers made tests (Pratiwi et al., 2019).

When teachers’ made test is used for a summative test, some trade-offs do exists. Alderson (2005) claims that the tests made by teachers have poor quality, and the results cannot be used to point out how students can improve. Many students have accessed the quality of teachers made tests in terms of validity, reliability (Furwana, 2019), and item facility (Madehang, 2018). In addition, Nurhalimah et al. (2019) found that teachers made tests have low discrimination index. In addition, a large percentage of the distractors used in multiple-choice tests were ineffective (Rohmah, 2019). A research study by Ing et al. (2015) found that teachers are not proficient in following the test construct. One factor affecting the quality of tests designed by teachers is the lack of knowledge regarding language assessment and testing (Ölmezer-Öztürk & Aydin, 2018). Most high school teachers only hold undergraduate degrees where the concept of assessment is offered in only one course. In-service professional development training tends to focus on practical test design; therefore, the quality of tests created depends much on teachers' English proficiency level, which was mostly low according to research by Triastuti (2020).

Paper-based TOEFL for classroom assessment

TOEFL is one of the most popularly used and accepted standardized English language tests for academic purposes. The official test is now delivered through the internet. However, many educational institutions conducted the paper-based version of the test, which is still widely accepted for local university admission and scholarship application requirements (Ananda, 2016). The paper-based version, most commonly known as Institutional TOEFL, consists of three sections, i.e. listening comprehension (50 items), structure & written expression (40 items), and reading comprehension (50 items). Listening comprehension tests understanding of short dialogues, extended-length dialogues, and short talks. The second section measures knowledge of English grammar through completion and error analysis tests. Finally, reading comprehension measures understanding of stated and implied meaning and vocabulary, in five 400-450-word academic texts (Cohen & Upton, 2007). The results of the test were converted into scaled scores by using Item Response Theory (IRT) with three-parameter logistics (3PL) model (Way & Reese, 1991), and the scores range between 310 and 667 (ETS, 2011).

As outlined previously, to have good quality, a test needs to satisfy the requirements of a good test in the test development process. The requirements of a good test are well met by standardized language tests such as TOEFL (Fulcher, 2010). Thus, using this test for a summative assessment in an EFL classroom guarantees that the result accurately represents the students' English proficiency level (Liskinasih & Lutviana, 2016), and it is also recommended by Bailey (2017). Using the result from the test, a teacher will be able to group students based on their language profile because TOEFL possesses a high discrimination index. Thus, they can select materials intended for the students with the corresponding level of language proficiency. However, even the shortest version of TOEFL, such as Institutional TOEFL, consists of three sections (Taufiq et al., 2018). It requires two hours to complete, with a 30-minute preparation, which is less feasible to be conducted in a classroom with this limitation. Therefore, a shorter version of this test is preferable. However, this shorter version needs to be able to be converted to a long version in order for the score to be meaningful for students, such as to classify students into relevant CEFR levels, ranging from A1 to C1 (Tannenbaum & Baron, 2011). This can be achieved by administering only one section and use the score to predict the scores for the other two sections.

Predicting scores in language testing

Many research studies have utilized prediction based on existing data to draw significant conclusions. One classic example is the conclusion that the learners who score high in vocabulary tests will obtain a good score for reading comprehension (Alavi & Akbarian, 2012; Mehrpour & Rahimi, 2010; Sen & Kuleli, 2015). Among others, prediction in language testing can be made through two methods of analysis, i.e. time series and regression analyses. Initially, time-series design is used to record data obtained in different periods of time (Best & Kahn, 2006; Blasco, 2015; Hatch & Lazaraton, 1991; Lyons & Doueck, 2010). When all time-series data are plotted into a line chart, the line can be used to predict future points, known as out-of-sample forecast (Chatfield, 2000), based on trends obtained in previous points (Mendenhall et al., 2013). The application of time-series is very common among language test developers such as Educational Testing Service, mostly to develop automated rating systems (Ramanarayanan et al., 2015).

The other procedure of making a prediction is regression analysis, which is "an extension of correlational analysis" (Dancey & Reidy, 2011). It is one of the statistical procedures of predicting a variable from another variable (Mackey & Gass, 2005). By using regression analysis, it is also possible to predict a variable based on multiple variables, termed as multiple regression (Kothari, 2004). In fact, multiple regression is claimed as the most commonly used model in empirical research studies (Kelley & Bolin, 2013). In language testing research, regression analysis has been very widely used such as in Hambleton et al. (1991) for calculation of a model in Item Response Theory. DeMauro (1992) used regression analysis to predicts scores among three standardized language tests, i.e. Test of Spoken English (TSE), Test of Written English (TWE), and Test of English as a Foreign Language (TOEFL). Zechner et al. (2007) estimated speaking scores between the automated scoring system and human rater in an internet-based TOEFL.

Research method

Data and source of data

This research was a quantitative research study where the data were numerical data, collected from the results of the Institutional Test of English as a Foreign Language (TOEFL ITP) conducted by the Language Center of Universitas Syiah Kuala, Indonesia, in conjunction with the English Testing Service (ETS) represented by the Indonesian International Education Foundation (IIEF). The scores were collected between 2007 and 2011 because the raw scores were no longer disclosed by the IIEF for the later dates. The number of scores collected was 2,030 scores as the training data sample to construct a linear regression model, obtained by 1,150 females (56.65%) and 880 males (43.35%) between 17 and 50 years old. The scores range from 417 and 653. In addition, the number of scores used for testing the models in this study was 102 scores.

Data analysis

To obtain the models of prediction for the total scores based on any sub-score, the data analysis was divided into two primary parts. The first part is a model assumption checking based on the training data sample. Bowerman et al. (2005, pp. 97–98) stated that those assumptions include (1) residual values of the model that must be independent of the value of X and normally distributed with the mean equal to zero and constant variances and (2) no pattern of positive (negative) residual terms is followed by other positive (negative) residual terms.

The second part of the analysis consists of the hypothesis testing of the regression model. There are two types of hypothesis testing in the regression model, i.e. model testing using F test and parameters testing using t-test. Besides the hypothesis testing, we also used a coefficient of determination (adjusted R^2) to measure how good the model was. A model with a higher value of adjusted R^2 was considered a better model. A linear regression model was created for each score where TOEFL ITP score as the response variable Y and the subtest scores (listening comprehension, structure & written expression, or reading comprehension) as the predictor variable X. Accordingly, there are three different linear regression models generated in this study.

Findings and discussion

Descriptive statistics

To evaluate the relationship between listening comprehension, structure & written expression, and reading comprehension skills, we examined 2,030 TOEFL scores as a training dataset. Table 1 shows the summary statistics of those data indicated by the minimum, median, mean, and maximum scores.

Table 1. Summary Statistics

Statistics	Listening	Structure	Reading	TOEFL Score
Minimum	9.00	5.00	7.00	417.00
Median	27.00	21.00	25.00	480.00
Mean	27.10	21.18	25.75	483.40
Maximum	50.00	38.00	49.00	653.00

Overall, Table 1 shows that the average score was 483, i.e. 3 points above the median. The minimum and maximum scores were 417 and 653 respectively. In this study, we restricted our minimum TOEFL scores to 417 as suggested by Mustafa and Anwar (2018), who show that the TOEFL score of 417 is considered as the lowest meaningful PBT TOEFL score for English proficiency assessment.

Distribution of training data sample

To visualize the distribution of score for each TOEFL ITP sub-test scores, Figure 1 shows the boxplot of the raw scores (true score) for listening comprehension (Figure 1a), structure & written expression (Figure 1b), and reading comprehension (Figure 1c) for each TOEFL score (scaled score). The minimum and maximum values were displayed as the bottom and upper sides of the whisker, the first and third quartiles were displayed as the bottom and upper parts of the box, and the median was displayed as the horizontal line in the box. The wider the size of the whiskers is, the wider range the data contain.

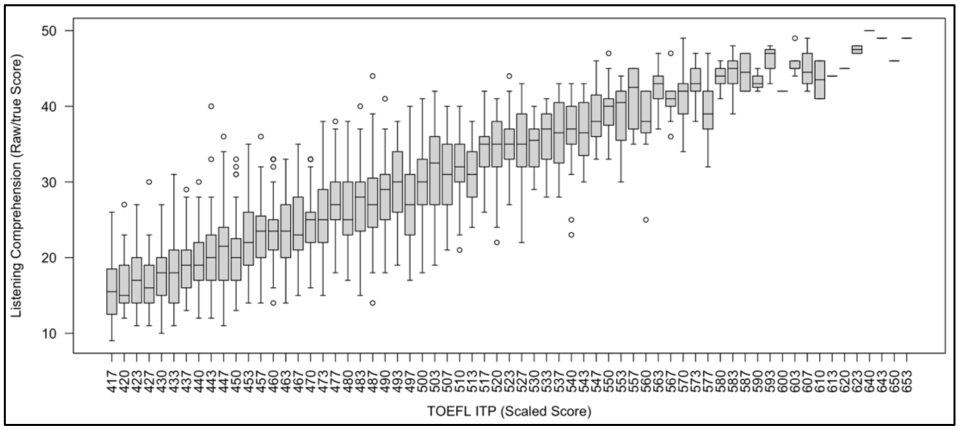


Figure 1a. Distribution of Training Data Sample for Listening Comprehension

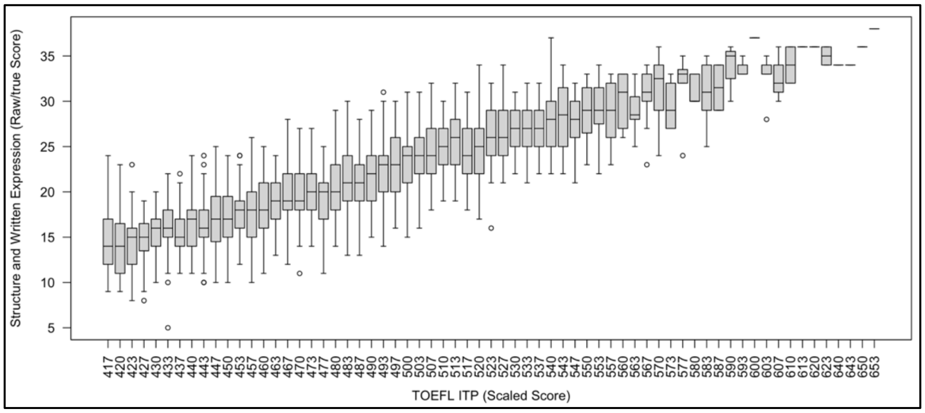


Figure 1b. Distribution of Training Data Sample for Structure & Written Expression

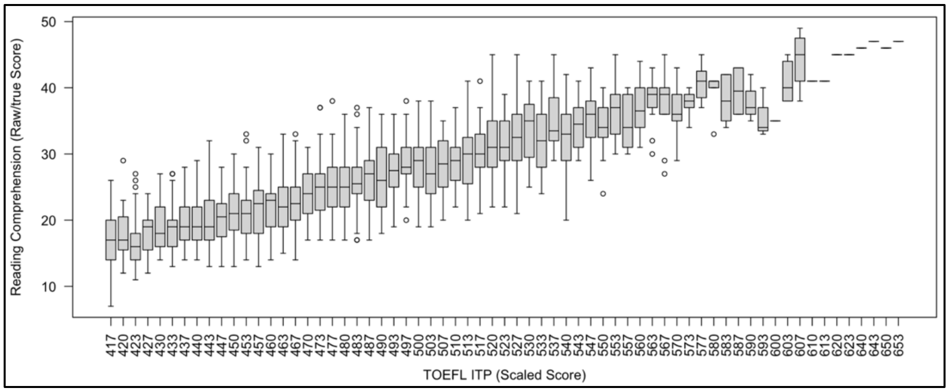


Figure 1c. Distribution of Training Data Sample for Structure & Written Expression

The three sub-figures in Figure 1 above show almost similar behavior, where lower TOEFL scores tended to have wider whiskers, and higher TOEFL scores tended to have narrower whiskers. A wider whisker indicates that the raw score for the respective section was more diverse for those TOEFL scores than the narrower one. In other words, test takers with lower TOEFL scores might have more variations in the raw scores in each section of the test than those with higher TOEFL scores. Another important piece of information shown in the boxplot is the extreme values indicated by the points (labeled by an empty circle) under the minimum or above the maximum values (whiskers). Figure 1 also shows that each section had several extreme values located above and under the whiskers. The listening comprehension section had more extreme values than other sections, which means that some test takers with similar TOEFL scores might have higher raw scores in the listening section than other test takers with the same TOEFL scores.

The linear regression model of the TOEFL ITP score

We evaluated the relationship between TOEFL scores as the response variable and its predictor variables to predict the TOEFL scores. The relationship was evaluated through a simple linear regression model with the following formula:

$$Y = \beta_0 + \beta_1 X + e \quad (1)$$

In formula 1, Y is the TOEFL score as the response variable, and X is listening comprehension, structure & written expression or reading comprehension as the predictor variable. The parameter model, β_0 and β_1 are called the intercept and slope of the model respectively. The error of the model is written as e, and it represents the difference between the response variable (Y) and the estimate of the response variable ($\hat{Y}$).

In this study, the intercept parameter counts for a TOEFL score when the predictor variable is assumed to be zero. In other words, it is the average TOEFL score when the effect of the predictor variable is not counted. However, the intercept in this study does not have any significant meaning because, based on ETS (2011, p. 14), the lowest TOEFL score is 310. Meanwhile, the slope parameter counts for the amount of increase or decrease in the TOEFL score when the predictor variable increases by one point. Those parameters are evaluated partially through the test statistic *t*. Furthermore, an *F* test can be used to test the significance of the regression relationship between the predictors and response variables (regression model).

Another important statistics in the linear regression analysis are correlation coefficient (r) and adjusted R^2. The correlation coefficient measures the strength of the linear relationship between response and predictor variables. While the adjusted R^2 is the proportion of total variation on the response variable explained by the linear regression model. Table 2 summarizes the result of the linear regression models for each predictor variable using the training data sample, while Table 3 and Figure 2 show the test of normality assumptions and scatterplots of the residuals respectively.

Table 2. Linear Regression Model Based on Training Data Sample

Model	Predictor variable	Parameter	Estimate	*t* value	*p-value*
1	Listening comprehension[a]	Intercept	373.07	213.02	< 0.001
		Slope	4.07	65.78	< 0.001
2	Structure & written expression[b]	Intercept	357.14	171.81	< 0.001
		Slope	5.96	62.73	< 0.001
3	Reading comprehension[c]	Intercept	364.19	177.42	< 0.001
		Slope	4.63	60.15	< 0.001

[a] *F-value* = 4326 (< 0.001), r = 0.8251, adjusted R^2 = 0.6807
[b] *F-value* = 3936 (< 0.001), r = 0.8124, adjusted R^2 = 0.6598
[c] *F-value* = 3617 (< 0.001), r = 0.8005, adjusted R^2 = 0.6406

Table 3. Residual Assumptions

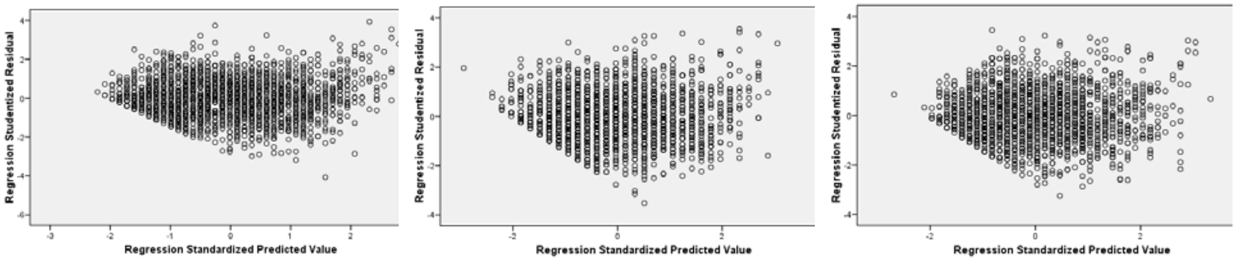
Response variable	Predictor Variables	Mean	Variance	Kolmogorov-Smirnov Z	*p-value*
TOEFL score	Listening comp.	0.000	22.798^2	0.956	0.320
	Structure & writ. exp.	0.000	23.533^2	1.340	0.055
	Reading comp.	0.000	24.187^2	1.312	0.064

Further, the linear regression models in Table 2 can be written as follow:

Model 1: $\text{TOEFL score (Y)} = 373.07 + 4.07 * \text{Listening Comprehension } (X_1)$ (2)

Model 2: $\text{TOEFL score (Y)} = 357.14 + 5.96 * \text{Structure \& Written Exp. } (X_2)$ (3)

Model 3: $\text{TOEFL score (Y)} = 364.19 + 4.63 * \text{Reading Comprehension } (X_3)$ (4)



(a) (b) (c)

Figure 2. Scatterplots of The Residuals

Model 1 is for the simple linear regression between TOEFL score as the response variable and listening comprehension sub-score as the predictor variable, and the estimated parameters of intercept and slope are 373.07 and 4.07 respectively. The slope of the parameter indicates that the TOEFL score increases by 4.07 points for each increase in raw score in the listening comprehension section. Both parameters were statistically significant in the test statistic t and F test (p<0.001). The TOEFL score and listening comprehension sub-score had a relationship of around 82.51%. The adjusted R^2 of model 1 was 0.6807, which suggests that 68.07% of the total variation on the TOEFL score was explained by the listening comprehension sub-score while the rest was explained by other variables that were not included in this study. Table 3 shows that the residuals of the model followed the normality distribution with zero mean and variance of 22.798^2 through the Kolmogorov-Smirnov test (p>0.05). Scatterplot of regression standardized residual versus regression standardized predicted value (Figure 2 (a)) shows there was no clear pattern that a positive (or negative) residual term was followed by other positive (or negative) residual terms. Accordingly, we can conclude that model 1 had fulfilled the required assumptions.

For model 2, the TOEFL score increases by around 6 points for each point increase in the structure & written expression section. Both parameters, i.e. the intercept and slope, were statistically significant. The TOEFL score had a correlation with the structure & written expression sub-score of 81.24%, and the structure sub-score explained 65.98% variability of the TOEFL score. The last model shows that the TOEFL score and the reading comprehension sub-score had a correlation coefficient of 80.05%. Based on the slope, the score increases by 4.63 points for each increase in the raw score of the reading comprehension section. The intercept and slope parameters of model 3 were also statistically significant. In addition, 64.06% of TOEFL score variability was explained by the reading comprehension sub-score. The residuals of the model 2 and 3 also met all required assumptions as did model 1.

The respective linear regression models are presented in Figure 3. The regression line (red) in the figure was constructed by using the parameter model (intercept and slope) for each paired data point (X, Y). The Y-axis represents the TOEFL score, while the X-axis represents the raw score (number of correct answers) in the listening comprehension, structure & written expression, and reading comprehension sections respectively for each sequential figure. The regression line in each figure tended to increase linearly with different intercepts. Those lines represent the predicted TOEFL score for each correct item in the test section of the training data sample.

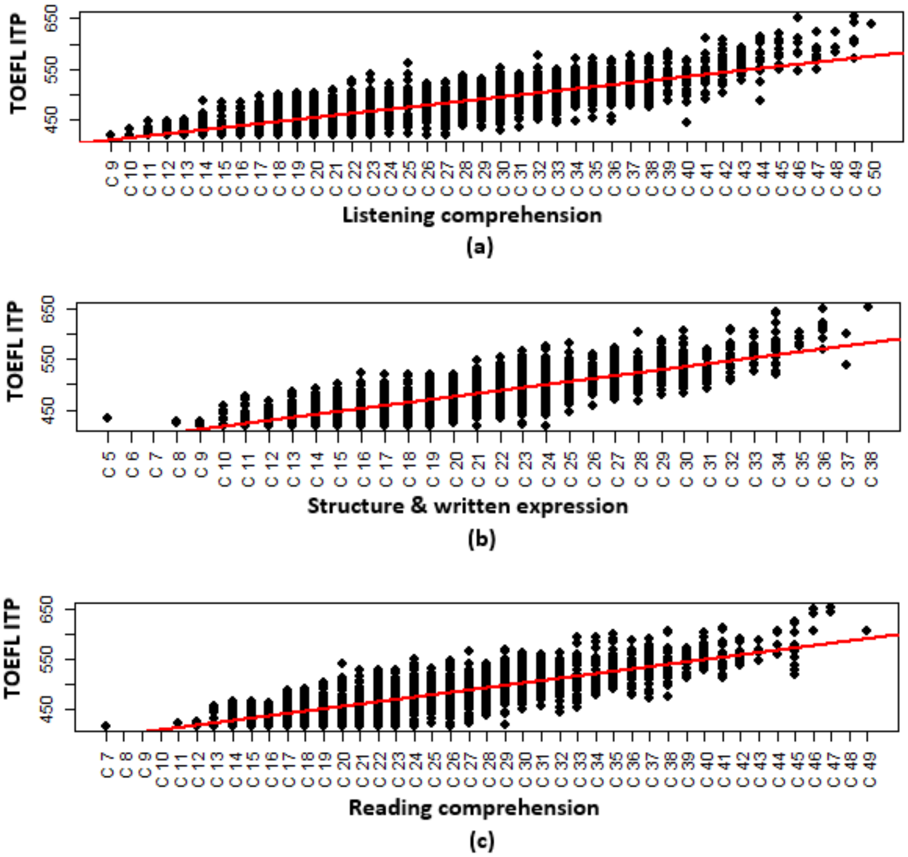


Figure 3. Linear Regression Plots

However, since the last possible numbers based on the TOEFL scoring system are 0, 3, and 7 (e.g. 470, 473, 477, …), those predicted TOEFL scores must be rounded to 0, 3, and 7 by using the following criteria:

- Rounded to 0 for last number 9, 0, and 1
- Rounded to 3 for last number 2, 3, and 4
- Rounded to 7 for last number 5, 6, 7, and 8

For the score ending in 9, in addition to rounding it to 0, one point is also added to the middle score (e.g. 479 into 480, etc.)

Using this transformation, the predicted TOEFL scores in this study follow the TOEFL scoring system. The accuracy of the prediction of the linear regression model could be measured by Mean Square Error (MSE) criteria using the formula:

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} \quad (5)$$

Where: y_i is true TOEFL score for the – i^{th} sample
$\hat{y}_i$ is predicted TOEFL score for the – i^{th} sample
n is the total sample size

MSE measures the average square error of the model by comparing the true TOEFL score with the predicted counterpart based on the linear regression model. The smaller MSE is, the better linear regression model performs. The MSEs of those three linear regression models applied to the training dataset are shown in Table 4.

Table 4. MSE of Training Data Sample

Data	MSE		
	Listening comprehension	Structure & written exp.	Reading comprehension
Training data sample (2,030 scores)	519.76	553.62	587.05

Table 4 shows that the model with the smallest MSE (519.76) is the linear regression model with the listening comprehension sub-score as the predictor variable (model 1). This suggests that model 1 performs better in predicting the TOEFL score than other models using the training data sample.

Prediction of paper-based TOEFL score on the testing dataset

The linear regression model constructed by using a training data sample can be used to predict the TOEFL score for a given new dataset of raw scores in each section of the TOEFL test. We used the model in equations (2), (3) and (4) to predict the TOEFL score for the new dataset by entering the raw scores of the listening comprehension, structure & written expression, and reading comprehension sections in respective equations. The predicted TOEFL score was then rounded based on the TOEFL scoring system using similar criteria as in the training dataset. The testing data consisted of the raw scores in all sections and also the respective TOEFL score. We then performed the linear regression model in the equation (2), (3) and (4) to predict the TOEFL score for that testing data sample. The raw scores in each section, the predicted TOEFL score, and the true TOEFL score of the first four and the last two testing data sample, are presented in Table 5.

Table 5. Prediction of TOEFL Score

Sample	Correct items in the Listening section	Predicted TOEFL Score	Correct items in the Structure section	Predicted TOEFL Score	Correct items in the Reading section	Predicted TOEFL Score	True TOEFL Score
1	21	460	20	477	25	480	467
2	27	483	21	483	21	460	460
3	18	447	22	487	31	507	467
4	19	450	23	493	30	503	470
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
101	13	427	14	440	24	477	437
102	14	430	17	457	24	477	450

A complete TOEFL scores prediction of those testing data samples (Table 5) can be observed in Figure 4 along with its true TOEFL score. The predicted TOEFL score with the listening comprehension sub-score (model 1) is represented by a red line, predicted TOEFL score with structure & written expression sub-score (model 2) is represented by the blue line. In addition, the yellow line represents model 3 (reading comprehension sub-score), while the black line shows the true TOEFL score.

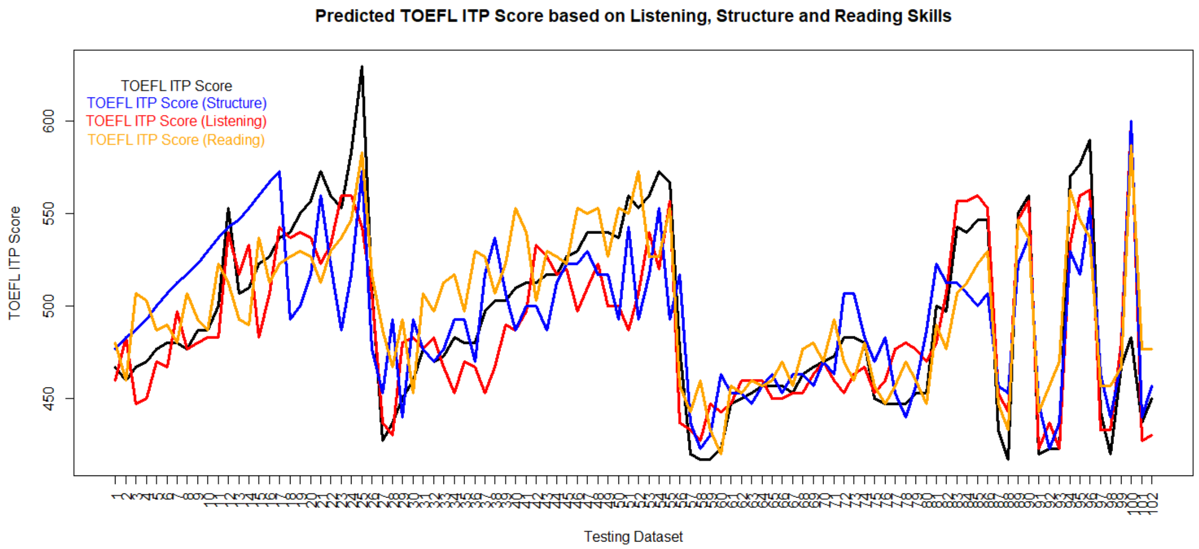


Figure 4. Predicted TOEFL Score for The Testing Data Sample

Figure 3 shows that the predicted TOEFL scores of each model seem to follow the true TOEFL score with different trends and fluctuation. As in the training data sample, the predicted correctness of the regression model applied to the testing data sample can be measured by MSE as in equation (5). The MSEs of those three linear regression models applied to the testing data sample are shown in Table 6.

Table 6. MSE of Testing Data Sample

Data	MSE		
	Listening comprehension	Structure & written exp.	Reading comprehension
Testing data sample (102 scores)	670.28	1002.43	710.04

As the training data sample, the prediction model with the smallest MSE was the linear regression model with listening comprehension sub-score as the predictor variable (model 1) with the MSE of 670.28. However, in the testing data sample, model 3 with reading comprehension sub-score as the predictor variable performed better than the model with structure & written expression sub-score to predict the TOEFL score for a given sample with the MSE of 710.04. As in the training data sample, we can conclude that model 1 performed better in predicting the TOEFL score than the other models for a given new dataset. However, model 3 is also recommended due to its low MSE value for the testing data sample. Furthermore, the MSE in the testing data sample was greater than the training data sample since the model was constructed based on the training data sample.

Discussion

This study aims to find the best model to predict English proficiency level represented by TOEFL scores using only one subtest out of three, i.e. listening comprehension, structure & written expression, and reading comprehension. One model was generated for each subtest using simple linear regression analysis, resulting in three models. The model was evaluated using Mean Squared Error (MSE), where lower the MSE means less error in the model prediction. Since the model involving listening comprehension sub-test as the independent variable produced the lowest error

(MSE = 670), which is close to the model with reading comprehension as the independent variable (MSE = 710). The MSE for the other sub-test was extremely higher compared to the earlier models. Therefore, the research results conclude that either the listening comprehension subtest or reading comprehension subtest can be used to predict the total score when it is not possible to conduct TOEFL with all subtests.

The listening comprehension subtest has been found to be the most accurate predictor for TOEFL because this subtest shows latent language skills. This result is motivated because listening skills are only developed when students' English proficiency has significantly improved (Graham, 2011, p. 113). In addition, research conducted by Wang and Treffers-Daller (2017) shows that progress in listening skills is a result of the improvement in general English language proficiency. This result can be treated as a warning in English language classes where listening comprehension is not included in the English language test due to difficulty of the testing procedure and test material development. This result also implies that listening instruction does not provide adequate impact on listening ability because the result of the instruction is mediated by other factors, i.e. general English language proficiency and vocabulary mastery (Wang & Treffers-Daller, 2017). Therefore, Rost (2014) suggests listening instruction uses a bottom-up process, where students are first facilitated to deal with easy material to develop their linguistic knowledge, one of which is vocabulary. This belief is also shared among professionally-trained English teachers in China (Li & Renandya, 2012).

The reading comprehension subtest was almost as accurate TOEFL score predictor as the listening comprehension counterpart. This result is also anticipated because reading, together with listening, is what facilitates language acquisition (VanPatten, 2015). Reading comprehension is correlated to vocabulary, which is correlated to overall English proficiency to a large extent (Miralpeix & Muñoz, 2018). Therefore, it is expected that the reading comprehension subtest can be used to accurately predict overall TOEFL score. In addition, reading is one of the channels for language input, leading to language acquisition in general. In fact, reading is the key agent which facilitates learner’s input to academic English in English as a foreign language context (McQuillan, 2019), which is what TOEFL is intended to measure. This research result is also supported by the result of the study conducted by Droop and Verhoeven (2003), which shows that reading proficiency increases as the overall proficiency level improves. This explains why the reading comprehension subtest is an accurate predictor of overall TOEFL score, representing overall English proficiency level.

Structure & written expression subtest measures test takers’ “knowledge of structural and grammatical elements of standard written English” (Educational Testing Service, 2013, p. 2). This subtest is not included in language skills because grammar can be learned as knowledge, which is different from acquisition because knowing a grammatical concept does not guarantee that a learner can use the concept in a conversation (Gass & Selinker, 2008). Applying perceived knowledge in a productive language skill such as speaking and writing is a complex proses, termed input processing (VanPatten, 2015). Structure & written expression support language production, but it is not by itself a language skill, and it does not contribute much to the overall language proficiency (Iwashita, 2018). A study by Kusumawardani and Mardiyani (2018) concludes that the correlation between grammar and productive language is weak.

As an implication for language testing, EFL teachers might now estimate their students' TOEFL score without the need to have them sit the complete version of the TOEFL test that consists of three sections, i.e. listening comprehension, structure & written expression, and reading comprehension sections. Instead, they can test only one of those sections to estimate their students'

TOEFL scores. Further, the listening comprehension sub-score could predict a better TOEFL score than could other skills (structure & written expression and reading comprehension sections) based on the MSE of both training and testing data. These regression models in the equation (2), (3), and (4) can be used by students, teachers, lecturers, managers or other stakeholders who need to know the estimation of TOEFL score for their respective purpose. In terms of pedagogical implication, the study has revealed that both listening comprehension and reading comprehension are significant predictors for overall English language proficiency, and both subtests are very dependent on vocabulary. Since vocabulary is strongly correlated to the overall level of language performance (Iwashita, 2018), teachers should integrate more indirect vocabulary instruction in English language classrooms.

Conclusion

The purpose of this research was to generate prediction models for paper-based TOEFL based on one sub-score. The research results show that it is possible to generate the prediction models by using linear regression analysis based on 2,030 previously obtained scores normally distributed at $p > 0.05$. The intercepts for the prediction models are 373.03 for listening comprehension, 357.14 for structure & written expression, and 364.19 for reading comprehension. Meanwhile, the slopes are 4.07 for listening comprehension, 5.96 for structure & written expression, and 4.63 for reading comprehension. Those intercepts and slopes are significant at $p < 0.001$. After applying the models to a testing data sample of 102 scores, we obtained the Mean Square Error (MSE) of 670 for listening comprehension, 1,002 for structure & written expression, and 710 for reading comprehension. These results suggest that listening comprehension and reading comprehension subtests are more accurate in predicting overal TOEFL scores than structure & written expression subtest.

Declaration of conflicting interest
The authors declare that there is no conflict of interest in this work.

Funding acknowledgement
The authors received no specific funding for this work.

References

Alavi, S. M., & Akbarian, I. (2012). The role of vocabulary size in predicting performance on TOEFL reading item types. *System*, *40*(3), 376–385. https://doi.org/10.1016/j.system.2012.07.002

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. Continuum.

Ananda, R. (2016). Problems with section two ITP TOEFL test. *Studies in English Language and Education*, *3*(1), 37–51. https://doi.org/10.17969/siele.v3i1.3387

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Bailey, A. L. (2017). Assessing the language of young learners. In E. Shohamy, L. G. Or, & S. May (Eds.), *Language Testing and Assessment* (pp. 323–342). Springer International Publishing. https://doi.org/10.1007/978-3-319-02261-1_22

Best, J. W., & Kahn, J. V. (2006). *Research in education* (10th ed.). Pearson Education Inc.

Blasco, M. E. (2015). A cognitive linguistic analysis of the cooking domain and its implementation in the EFL classroom as a way of enhancing effective vocabulary teaching. *Procedia - Social and Behavioral Sciences*, *178*, 70–77. https://doi.org/10.1016/j.sbspro.2015.03.149

Bowerman, B. L., O'Connell, R. T., & Koehler, A. B. (2005). *Forecasting, time series, and regression: An applied approach*. Thomson Brooks/Cole. https://books.google.co.id/books?id=2Yc_AQAAIAAJ

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. Longman.

Brown, J. D. (1996). *Testing in language programs*. Prentice Hall Regents.

Chatfield, C. (2000). *Time-series forecasting*. Chapman & Hall/CRC.

Cohen, A. D., & Upton, T. A. (2007). `I want to go back to the text': Response strategies on the reading subtest of the new TOEFL(R). *Language Testing*, *24*(2), 209–250.

Coniam, D. (2009). Investigating the quality of teacher-produced tests for EFL students and the effects of training in test development principles and practices on improving test quality. *System*, *37*(2), 226–242. https://doi.org/10.1016/j.system.2008.11.008

Dancey, C., & Reidy, J. (2011). Statistics without maths for psychology. In *Book*. http://books.google.com/books?hl=en&lr=&id=QjfQ0_DqyNQC&oi=fnd&pg=PR16&dq=Statistics+Without+Maths+for+Psychology&ots=5PBfHf-mB-&sig=XUC1_n2l4AVh3o_qgCh7wE8FmuY

DeMauro, G. (1992). Examination of the relationships among TSE, TWE and TOEFL scores. *Language Testing*, *9*(2), 149–161. https://doi.org/10.1177/026553229200900203

Douglas, D. (2010). *Understanding language testing* (B. Comrie & G. Corbett (eds.)). Routledge.

Droop, M., & Verhoeven, L. (2003). Language proficiency and reading ability in first- and second-language learners. *Reading Research Quarterly*, *38*(1), 78–103. https://doi.org/10.1598/RRQ.38.1.4

Educational Testing Service. (2013). *Official guide to TOEFL ITP test*. Educational Testing Service.

ETS. (2011). *Test and score data summary for TOEFL® Internet-based and paper-based tests: January 2010 - December 2010 test data*. https://www.ets.org/Media/Research/pdf/TOEFL-SUM-2010.pdf

Fulcher, G. (2010). *Practical language testing*. Hodder Education. https://doi.org/10.4324/9780203767399

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Routledge. https://doi.org/10.1177/026553229301000104

Furwana, D. (2019). Validity and reliability of teacher-made English summative test at second grade of Vocational High School 2 Palopo. *Language Circle: Journal of Language and Literature*, *13*(2). https://doi.org/10.15294/lc.v13i2.18967

Gass, S. M., & Selinker, L. (2008). *Second language acquisition: An introductory course* (2nd ed.). Routledge Taylor & Francis Group.

Graham, S. (2011). Self-efficacy and academic listening. *Journal of English for Academic Purposes*, *10*(2), 113–117. https://doi.org/10.1016/j.jeap.2011.04.001

Green, A. (2014). *Exploring language assessment and testing: Language in action*. Routledge Taylor & Francis Group. https://doi.org/10.4324/9781315889627

Hambleton, R. K., Swaminathan, H., & Rogers, D. J. (1991). *Fundamentals of item response theory*. Sage Publications.

Hatch, E., & Lazaraton, A. (1991). The research manual: Research design and statistics for applied linguistics. In *The Modern Language Journal*. Heinle & Heinle Publishers. https://doi.org/10.2307/327087

Henning, G. (1987). *A Guide to language testing: Development, evaluation and research*. Foreign Language Teaching and Research Press.

Ing, L. M., Musah, M. B., Al-Hudawi, S. H. V., Tahir, L. M., & Kamil, N. M. (2015). Validity of teacher-made assessment: A table of specification approach. *Asian Social Science*, *11*(5), 193–200. https://doi.org/10.5539/ass.v11n5p193

Iwashita, N. (2018). Grammar and language proficiency. In J. I. Liontas (Ed.), *The TESOL Encyclopedia of English Language Teaching* (pp. 1–7). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118784235.eelt0069

Kelley, K., & Bolin, J. H. (2013). Multiple regression. In T. Teo (Ed.), *Handbook of quantitative methods for educational research* (pp. 71–101). Sense Publishers.

Kothari, C. R. (2004). *Research methodology: Methods and techniques* (2nd Ed). New Age International (P) Ltd.

Kusumawardani, S. A., & Mardiyani, E. (2018). the Correlation Between English Grammar Competence and Speaking Fluency. *PROJECT (Professional Journal of English Education)*, *1*(6), 724–733. https://doi.org/10.22460/project.v1i6.p724-733

Li, W., & Renandya, W. A. (2012). Effective approaches to teaching listening: Chinese EFL teachers' perspectives. *Journal of Asia TEFL*, *9*(4), 79–111.

Liskinasih, A., & Lutviana, R. (2016). The validity evidence of TOEFL test as placement test. *Jurnal Ilmiah Bahasa Dan Sastra*, *3*(2), 173–180. https://doi.org/10.21067/jibs.v3i2.1513

Lyons, P., & Doueck, H. J. (2010). *The dissertation: From beginning to end*. Oxford University Press.

Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design*. Lawrence Erlbaum Associates. http://medcontent.metapress.com/index/A65RM03P4874243N.pdf%5Cnhttp://books.google.com/books?hl=en&lr=&id=b3CxLrJ_1pYC&oi=fnd&pg=PP1&dq=Second+Language+Research+Methodologie+and+Design&ots=GB2Lp7MNqy&sig=Hcm9uWbR6Zf27VYO2YlrfH85_0M

Madehang. (2018). The analysis of the English teacher-made tests based on the taxonomy of instructional bbjectives in the cognitive domain at the state senior secondary schools in Palopo. *Asian EFL Journal*, *50*(5), 221–227.

McQuillan, J. (2019). Where do we get our academic vocabulary? Comparing the efficiency of direct instruction and free voluntary reading. *Reading Matrix: An International Online Journal*, *19*(1), 129–138.

Mehrpour, S., & Rahimi, M. (2010). The impact of general and specific vocabulary knowledge on reading and listening comprehension: A case of Iranian EFL learners. *System*, *38*(2), 292–300. https://doi.org/https://doi.org/10.1016/j.system.2010.01.004

Mendenhall, W. I., Beaver, R. J., & Beaver, B. M. (2013). *Introduction to Probaility and Statistics*. https://doi.org/10.1017/CBO9781107415324.004

Miralpeix, I., & Muñoz, C. (2018). Receptive vocabulary size and its relationship to EFL language skills. *IRAL - International Review of Applied Linguistics in Language Teaching*, *56*(1), 1–24. https://doi.org/10.1515/iral-2017-0016

Mustafa, F., & Anwar, S. (2018). Distinguishing TOEFL score: What is the lowest score considered a TOEFL score? *Pertanika Journal of Social Sciences and Humanities*, 26(3), 1995–2008.

Nurhalimah, N., Fahriany, F., & Dadan, D. (2019). Determining the quality of English teacher-made test: How excellent is excellent? Indonesia. *Indonesiann EFL Journal: Journal of ELT,* *Linguistics, and Literature, 5*(1), 24–38.

Ölmezer-Öztürk, E., & Aydin, B. (2018). Investigating language assessment knowledge of EFL teachers. *Hacettepe University Journal of Education*, *34*(3), 1–19. https://doi.org/10.16986/HUJE.2018043465

Pratiwi, N. P. W., Dewi, N. L. P. E. S., & Paramartha, A. A. G. Y. (2019). The reflection of HOTS in EFL teachers' summative assessment. *Journal of Education Research and Evaluation*, *3*(3), 127–133. https://doi.org/10.23887/jere.v3i3.21853

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education*, *4*(1), 1–11. https://doi.org/10.1080/2331186X.2017.1301013

Ramanarayanan, V., Chen, L., Leong, C. W., Feng, G., & Suendermann-Oeft, D. (2015). An analysis of time-aggregated and time-series features for scoring different aspects of multimodal presentation data. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 1373–1377.

Rohmah, N. (2019). Validity and reliability study on teacher-made assessment for English mid-term examination. *Proceedings of the Eleventh Conference on Applied Linguistics (CONAPLIN 2018)*, *254*, 107–110. https://doi.org/10.2991/conaplin-18.2019.236

Rost, M. (2014). Exploring EFL Fluency in Asia. In T. Muller, J. Adamson, P. S. Brown, & S. Herder (Eds.), *Exploring EFL Fluency in Asia*. Palgrave Macmillan UK. https://doi.org/10.1057/9781137449405

Sen, Y., & Kuleli, M. (2015). The Effect of Vocabulary Size and Vocabulary Depth on Reading in EFL Context. *Procedia - Social and Behavioral Sciences*, *199*, 555–562. https://doi.org/10.1016/j.sbspro.2015.07.546

Shin, Y. K., & Kim, Y. J. (2017). Using lexical bundles to teach articles to L2 English learners of different proficiencies. *System*, *69*, 79–91. https://doi.org/10.1016/j.system.2017.08.002

Sulistyo, T., Eltris, K. P. N., Mafulah, S., Budianto, S., Saiful, S., & Heriyawati, D. F. (2020). Portfolio assessment: Learning outcomes and students' attitudes. *Studies in English Language and Education*, *7*(1), 141–153. https://doi.org/10.24815/siele.v7i1.15169

Tannenbaum, R. J., & Baron, P. A. (2011). *Mapping the TOEFL® ITP tests onto the Common European Framework of Reference*.

Taufiq, W., Santoso, D. R., & Fediyanto, N. (2018). Critical analysis on TOEFL ITP as a language assessment. *Advances in Social Science, Education and Humanities Research*, *125*, 226–229. https://doi.org/10.2991/icigr-17.2018.55

Triastuti, A. (2020). Assessing english pre-service teachers' knowledge base of teaching: Linking knowledge and self-portrayal. *TEFLIN Journal*, *31*(1), 108–138. https://doi.org/10.15639/teflinjournal.v31i1/108-138

VanPatten, B. (2015). Input processing in adult SLA. In B. VanPatten & J. Williams (Eds.), *Theories in Second Language Acquisition: An Introduction* (2nd ed., pp. 113–134). Routledge Taylor & Francis Group. https://doi.org/10.4324/9780203628942-12

Wang, Y., & Treffers-Daller, J. (2017). Explaining listening comprehension among L2 learners of English: The contribution of general language proficiency, vocabulary knowledge and metacognitive awareness. *System*, *65*, 139–150. https://doi.org/10.1016/j.system.2016.12.013

Way, W. D., & Reese, C. M. (1991). An investigation of the use of simplified IRT models for scaling and equating the TOEFL test. *ETS Research Report Series*.

Williams, C. H. (2017). *Teaching English in East Asia: A teacher's guide to Chinese, Japanese,* *and Korean learners*. Springer Nature Singapore Pte Ltd.

Zechner, K., Higgins, D., & Xi, X. (2007). SpeechRaterTM: A construct-driven approach to scoring spontaneous non-native speech. *Proceedings of the 2007 Workshop of the International Speech Communication Association (ISCA) Special Interest Group on Speech and Language Technology in Education (SLaTE2007)*, 128–131.