INDONESIAN JOURNAL OF APPLIED LINGUISTICS
Vol. 10 No. 1, September 2020, pp. 491-501

Available online at:
https://ejournal.upi.edu/index.php/IJAL/article/view/28591

https://doi.org/10.17509/ijal.v10i2.28591

The range of TOEFL scores predicted by TOEP

Suwarsih Madya, Heri Retnawati, Ari Purnawan, Nur Hidayanto P. S. Putro*, and Kartianom
Department of English Education, Faculty of Languages and Arts, Universitas Negeri Yogyakarta,
Jl. Colombo No. 1 Yogyakarta 55281, Indonesia

ABSTRACT

The Indonesian Testing Service Centre (ITSC) has developed an online standardized test called TOEP (Test of English Proficiency) as a fresh alternative for measuring the test takers' listening and reading proficiency. To ensure its quality, the TOEP scores need to be validated against the scores obtained from another established standardized test, in this case the ITP-TOEFL. This study aimed at finding out to what extent the range of scores which are measured by TOEP can predict the scores obtained from ITP-TOEFL. A quantitative approach was applied in this study, focusing on the analysis of scores obtained by 1,048 people taking TOEP in 2016, 2017, and 2019 and 383 testees had taken both TOEP and ITP-TOEFL. A regression analysis was conducted to establish the prediction equation of TOEP to ITP-TOEFL. The range of scores of proficiency measured through TOEP was estimated using the advanced item response theory, especially the information function value. The results of analysis show that TOEP can predict test takers' English proficiency in the range of minimum 310 and maximum 656.34 at the ITP-TOEFL scales. It can be concluded that TOEP has a good predictive validity to ITP-TOEFL.

Keywords: Interval of proficiency; prediction score; TOEFL; TOEP

First Received:	**Revised:**	**Accepted:**
17 July 2019	28 October 2019	24 May 2020
Final Proof Received:		**Published:**
29 June 2020		30 September 2020

How to cite (in APA style):
Madya, S., Retnawati, H., Purnawan, A., Putro, N. H. P. S., & Kartianom. (2020). The range of TOEFL scores predicted by TOEP. *Indonesian Journal of Applied Linguistics, 10*(2), 491-501. https://doi.org/10.17509/ijal.v10i2.28591

INTRODUCTION

The measurement of English proficiency has been an immense issue to a significant number of people particularly because the result of the test they took is a reflection of their communication skill –a necessary key to effectively survive in the 21st century and global world (Ahn, 2015; Mohammadi & Enayati, 2018; Saito, 2019). The need for good scores in proficiency tests increases significantly in the last decades as the result of the growing requirements for getting a job or continuing education. Only those with a high level of English proficiency can get good jobs in competitive environments and are more preferable in the work world (Lee, 2006; Simion, 2012). To measure one's English proficiency is then a necessity, certainly through a valid and reliable standardized test whose results are recognized world-wide.

In relation with the above concern, the existence of institutions providing mesurements and assesment of English competences is of prominence. People's English proficiency has so far been measured through standardized tests developed by native speakers of English, for example, the TOEFL, TOEIC, and IELTS tests. The development of such tests is very costly to ensure the standardization of such tests, which is reached through a series of activities to ensure that the test fulfills the criteria of a good test. A good test has to meet the validity criterion. Seen from the development process, the above said tests have good content validity or have been proven to be valid in terms of the content (Sawaki & Sinharay, 2018). As these tests have fulfilled the criteria of good tests in terms of content, the results are highly accurate in describing the level of test takers' English proficiency.

The high cost and expensive process of the test development result in a number of difficuties, one of which is in the fee for taking the tests. The tests become expensive, and some cannot afford them. For self evaluation purposes, more affordable

* Corresponding Author
Email: noersabar@yahoo.com

options are not always available. With difference in currency rates among countries, test takers from countries with weaker rates will have to spend a large sum of money. Another problem often found in Indonesia and some other countries is the limited access to such tests. Those tests are available only in big cities, and taking the tests is financially disastrous: the test takers have to spend extra money for transportation and accommodation in addition to the high test fee. The long, tedious process of registration, test administration, and the issuance of the test results are other additional problems encountered by the test takers. The long delay resulted from the slow process of score and certificate issuance sometimes obstucts users, especially when they need quicker test results.

As a response to the the above drawbacks, to offer test takers with alternatives, and to reduce dependency of English testing on tests provided by foreign agencies, the Association for the Teaching of English as a Foreign Language in Indonesia (TEFLIN) and the Association of Indonesian Psychology have co-founded the Indonesian Centre for Testing Services, of which the main programs are to develop TPDA or Test of Basic Academic Potential and TOEP or Test of English Proficiency. TOEP is an affordable test developed to measure Indonesian test-takers' English proficiency.

Similar to TOEFL, TOEP has also been developed through a series of activities, from the formulation of the purposes of constructing the test, the formulation of indicators of test items, the construction of test items, expert validation involving experts of language testing and psychometrics, to try-outs and calibration using the item response theory. This test consists of 100 items, covering 50 listening test items and 50 reading test items. The test takers' English proficiency is represented through the scores, which range from 0 (non-user of English) to 100 (expert user). Seen from the process of its development, TOEP is standardized and has fulfilled the criteria of a good test in terms of content validity.

Retnawati (2016) confirms that TOEP possesses the good criterion validity. A test taker's TOEP score could be used to predict his or her TOEFL score. The TOEP criterion validity is concurrent validity. This type of validity tells to what extent the result estimates the ability of another measurement instrument taken in about the same time (Fernandes, 1984). To provide evidence for the validity, two instruments are needed to measure the same construct, one being the predictor, i.e. the instrument of which the criteria validity will be proven, and the other being the criterion, i.e. the standardized measurement instrument such as TOEFL

Related to the concurrent validity of TOEP, the results of this test can be used to predict the scores achieved by taking other standardized tests, such as TOEFL ITP or IBT. This process is also done by educators and test developers in several other countries. In Japan, it was conducted by utilizing the results of *English Language Teaching and Learning* (Saito, 2019). In Iran, the validation and prediction were carried-out by utilizing the Cloze test (Saeedi et al., 2011). In Vietnam, the C-test was utilized to predict the scores of TOEFL, TOEIC, and IELTS (Hiser & Ho, 2016). Those local test developers conducted validation processes for their self-developed tests and administered them locally in their own countries for their particular purposes. However, other parties, especially from other countries, interested in using their tests for validation precesses or assessment procedures have difficulties in accessing those tests.

In addition to utilizing the test scores to predict the scores obtained through other tests, the results of a test or a measurement instrument which has been proven to have criterion validity can also be converted into the scores of other standardized tests. For example, the TOEP scores are converted into the ITP or IBT TOEFL scores (Retnawati, 2016). However, it is important to find out the degree of accuracy of the results or scores obtained from the predictor measurement instrument such as TOEP in predicting or describing the test takers' ability in other standardized tests such as TOEFL. This can be conducted by estimating the test takers' ability with as low measurement error as possible (Desjardins & Bulut, 2017) and utilizing information function values to find out the range of valid and reliable ability (Retnawati, 2016).

The test takers' ability can be estimated through the classical test theory and item response theory (Hambleton & Swaminathan, 2013). The classical test theory is based on an additive model, i.e. the test takers' true scores being obtained by deducting the observed scores by measurement error. Meanwhile, the item response theory is based on the probabilistic model, i.e. the test takers respond to the items based on the levels of the underlying latent nature. The test takers' ability in the item response theory (IRT) is estimated by considering item parametres (item difficulty range, the item differing power, and pseudo guessing). The classical test theory focuses more on the observed scores (raw scores) than considering the item relation with the measured latent trait (Desjardins & Bulut, 2017). Accordingly, when the focus of study is on predicting the test takers' ability, the item response theory is more preferable.

There are four models of the item response theory that can be used to estimate the test takers' ability by dichotomous scoring (1/0), the one-parameter model or 1-PL model, the Rasch model, the two parameter or 2-PL model, and the three-parameter or 3-PL model (Desjardins & Bulut, 2017). In the 1-PL model and the Rasch model, the test takers' ability is estimated by considering the

Copyright © 2020, *authors*, e-ISSN: 2502-6747, p-ISSN: 2301-9468

parameter of item difficulty range. For the 3-PL model, the test takers' ability is estimated by considering the difficulty range parameter, item discriminating indices, and the pseudo guessing. For the 2-PL model, the test takers' ability is estimated by considering the difficulty range parameter and item discriminating indices, and for 1-PL, the test takers' ability is estimated by considering the difficulty range parameter. Of the four non-linear models, the 1-PL model and Rasch model are the simplest. Compared to the 1-PL model, the Rasch model is more used by researchers or academics in various disciplines to meet the needs for item analysis and estimating the test takers' ability. Mathematically, the Rasch model can be presented as follows (Rasch, 1960):

$$P\left(Y_{ij}=1 \mid \theta_j, b_i\right)=\frac{\exp\left(D\left(\theta_j-b_i\right)\right)}{1+\exp\left(D\left(\theta_j-b_i\right)\right)},$$

where θ_j is the test takers' ability, *j* (*j* = 1, 2, …, *J*), b_i is the level of item difficulty of item *i* (*i = 1, 2, …, I*), and D is the scaling constancy to place the logistic parameter model at the normal ogive model scale (when D = 1.7). Equation (1) states that the probability of the test takers in responding correctly to the test item *i* is the function of ability and the level of item difficulty. The item discriminating power parameter is not included in Equation (1) because the item discriminating power is regarded as 1.

Another concept which is also important in the item response theory is the information function value. Item information function gives information related to the test item contribution in revealing the latent trait (ability) measured through a test (Hambleton et al., 1991; Retnawati, 2014; Desjardins & Bulut, 2017). Mathematically, the item information function can be presented as Equation 2 as follows.

$$I_i(\theta)=\frac{\left[P_i'(\theta)\right]^2}{P_i(\theta)Q_i(\theta)},$$

where *i* is an item (*i* = 1, 2, …, n), I_i is the item information function *i*, $P_i(\theta)$ is the probability for the test takers with ability θ to respond to item *i* correctly, $P_i'(\theta)$ is the function transferred from $P_i(\theta)$ to θ, and $Q_i(\theta)$ is the probability for test takers with ability θ to respond to item *i* wrongly. Because items are regarded as locally independent (the IRT assumption), the item information function for all items can be summed up and test information function calculated (Desjardins & Bulut, 2017).

Test information function is the sum of the test item information function (Hambleton et al., 1991; Wu et al., 2016). Test information function will be high if the item information function is also high. Test information function can be used to compare two test sets and to find out the ability traits appropriate for the test (Desjardins & Bulut, 2017; Hambleton et al., 1991; Retnawati, 2016). Mathematically, the test information function can be presented in Equation 3 as follows (Desjardins & Bulut, 2017; Hambleton et al., 1991; Retnawati, 2014).

$$I(\theta)=\sum_{i=1}^{n} I_i(\theta),$$

where the values of the item parameter index and test takers' ability are the estimation results. Since they are the result of estimation, the truth is probabilistic and not free from measurement error.

The Standard Error of Measurement or SEM in IRT is closely related to the information function value. The information function value has a negative quadratic relationship with SEM, i.e. the greater the information function value, the smaller the SEM or vise versa (Hambleton et al., 1991). If the information function value is stated by $I(\theta)$ and the estimation value of SEM by $\hat{SEM}(\theta)$, the relationship between the two can be represented mathematically in the same ways as in Equation 4 as follows (Hambleton et al., 1991; Retnawati, 2014; Desjardins & Bulut, 2017).

$$\hat{SEM}(\theta)=\frac{1}{\sqrt{I(\theta)}}$$

Studies on the validity of English tests have been conducted before, but they were focused on finding evidence for the content validity, construct validity, and criterion validity of the tests concerned. The previous studies on the criterion validity of English test were more focused on the predictive validity types, i.e. relating the TOEFL scores to students' grade point average (GPA). Meanwhile, very few studies on the criterion validity especially the concurrent validity have been conducted, but limited to the correlation of the two tests investigated. It should be noted that the correlation of the two tests have used raw score of the results of the analysis using the classical test theory. The research on TOEP has been limited to finding evidence of the TOEP criterion validity by correlating TOEP scores to TOEFL scores. However, the question on how accurate the scores obtained from the predictor test (e.g. TOEP scores) can predict or describe the test takers' ability on other standardized tests (e.g. TOEFL scores) has not been researched and is worth investigating.

Copyright © 2020, *authors*, e-ISSN: 2502-6747, p-ISSN: 2301-9468

Based on the literature review presented above, the result of tests (e.g. TOEP) with good criterion validity can be used to predict the test takers' scores provided that the test takers' ability is estimated by using the item response theory. In addition, the results of using the item response theory are the information function value of the items and the information function value of the test which can be used to find out the range of reliable scores of the test so that the results can be converted into other standardized tests. In this way, the purpose of the present paper is to find out the range of scores measured through TOEP administered in 2016 and 2017.

METHOD

This study was an explorative, descriptive quantitative study to describe the interval of TOEFL scores predicted by TOEFL score employing correlation and regression analyses. The data were collected through analysis of documents of 1,148 people taking TOEP which includes Listening and Reading in 2016 (362 test takers), 2017 (384 test takers), and 2019 (402 test takers) and 383 people taking both TOEP and ITP-TOEFL. These test takers were randomly selected to ensure they represented test takers with low, middle, and high theta or latent ability. Six TOEP forms with 50 listening test items and 50 reading test items in each test form were administered in 2016, four in 2017 and four in 2019. The data from the 1,148 testees were analyzed using the correlation and regression analyses. This study also used the Rasch model with one Parameter, i.e. the item difficulty level to analyze the items and calculate the information function values and the conversion scores. This involved the following steps: (1) establishing the criterion validity of TOEP against TOEFL by using the scores of TOEP Listening and Reading to predict those of TOEFL Listening and Reading as to obtain the prediction equation of the TOEFL scores; (2) estimating the parameter of the level of item difficulty of ten test forms by using test takers' response through the Rasch model; (3) calculating the information function value and SEM of each form; (4) determining the ability range at the standard normal scale that can be measured well by each form using the information function value and SEM; (5) converting the standard normal scale of Listening and Reading in TOEP (the 0-50 scale): (6) using the TOEP scores of both Listening and Reading to predict the TOEFL scores; and (7) determining a range of scores measured well through TOEP into the TOEFL scores. This was followed by interpreting the results.

FINDINGS

As has been touched upon before, to find out the predictive validity of TOEP to TOEFL, data were collected by administering the TOEFL test to 383 people out of 1,148 TOEP test takers. These 383 people were randomly selected by considering their theta or ability.

The predictive validity was established by estimating the correlation patterns between the TOEP Listening, Reading, and the total scores and the TOEFL Listening, Reading and the total scores. The analysis was conducted by observing the scatter plot, estimating the correlation and contribution, estimating the prediction equation, and estimating the errors at the total score model (TOEP with TOEFL) and the TOEP Listening and Reading scores as the predictor of TOEFL. Since the equivalence of the test forms has been tested and the TOEP test forms are found equal, the scores obtained from the tests are treated as equal (Madya et al., 2019). The findings of each analysis are presented below.

Figure 1 indicates a linear relationship between the TOEP Listening scores and the TOEFL Listening scores. The higher the TOEP listening scores are, the higher the TOEFL Listening scores will be. This indicates a positive correlation between the two tests, which means that the TOEP Listening score is a good predictor for the TOEFL Listening score. The same case applies to the Reading scores. That is, the TOEP Reading score is the predictor for the TOEFL Reading score. These are illustrated in Figures 2 and 3.

If the TOEP Listening and Reading scores are used together to predict the TOEFL scores, the scatter plot shows that the two variables serve as the predictor for the TOEFL scores. The higher the TOEP Listening and Reading scores are, the higher the TOEFL scores will be. This is illustrated in Figure 4.

After obtaining the information on the TOEP scores as the predictor based on the graphic, the estimation was conducted on the correlation, contribution and prediction equation of the TOEP scores to the TOEFL scores. The results are presented in Table 1.

The results of analysis indicated that the TOEP Listening scores and the TOEFL Listening scores were positively-correlated, with 63.5% contribution. Meanwhile, the TOEP Reading scores and the TOEFL Reading scores were positively-correlated with the contribution of 68.8%. These two contributions fell into the medium category. A higher contribution (79,6%) was obtained when the TOEFL score was predicted by the TOEP total scores (Listening and Reading). The highest contribution of 80.1% was obtained if the TOEP Listening and Reading scores were used together to predict the TOEFL score.

Copyright © 2020, *authors*, e-ISSN: 2502-6747, p-ISSN: 2301-9468

Figure 1
The scatter plot the TOEP listening score against the TOEFL listening

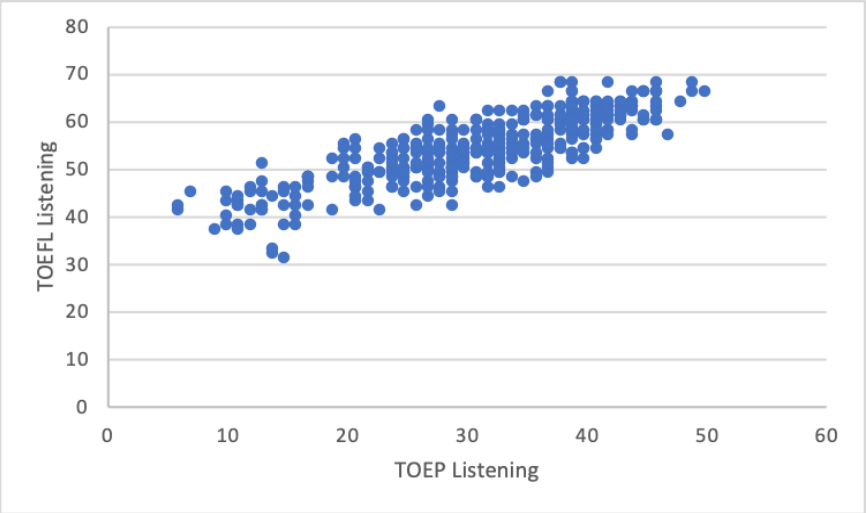


Figure 2
The scatter plot of TOEP reading scores against the TOEFL reading

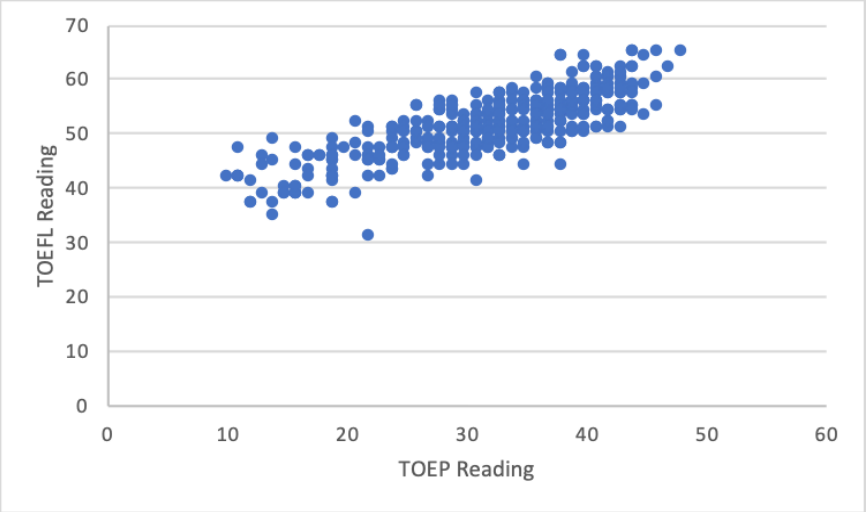


Figure 3
The scatter plot of TOEP scores against the TOEFL scores

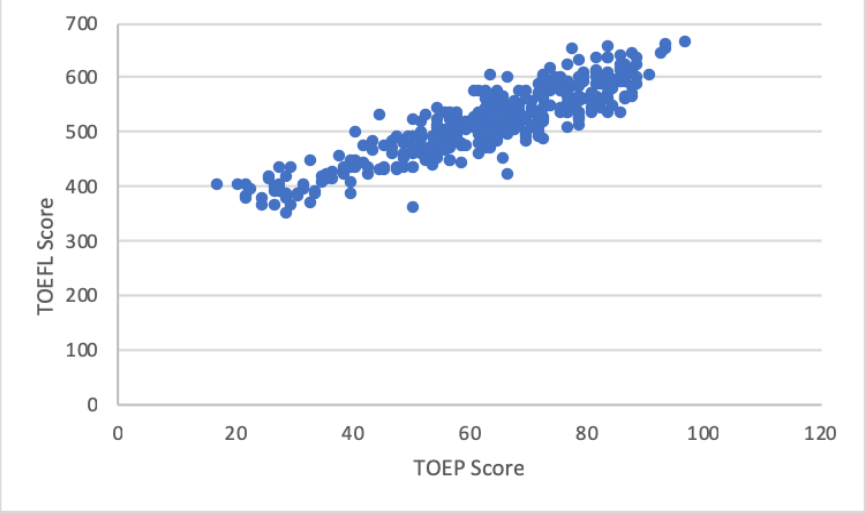


Figure 4
The scatter plot of listening and reading against the TOEFL scores

Copyright © 2020, *authors*, e-ISSN: 2502-6747, p-ISSN: 2301-9468

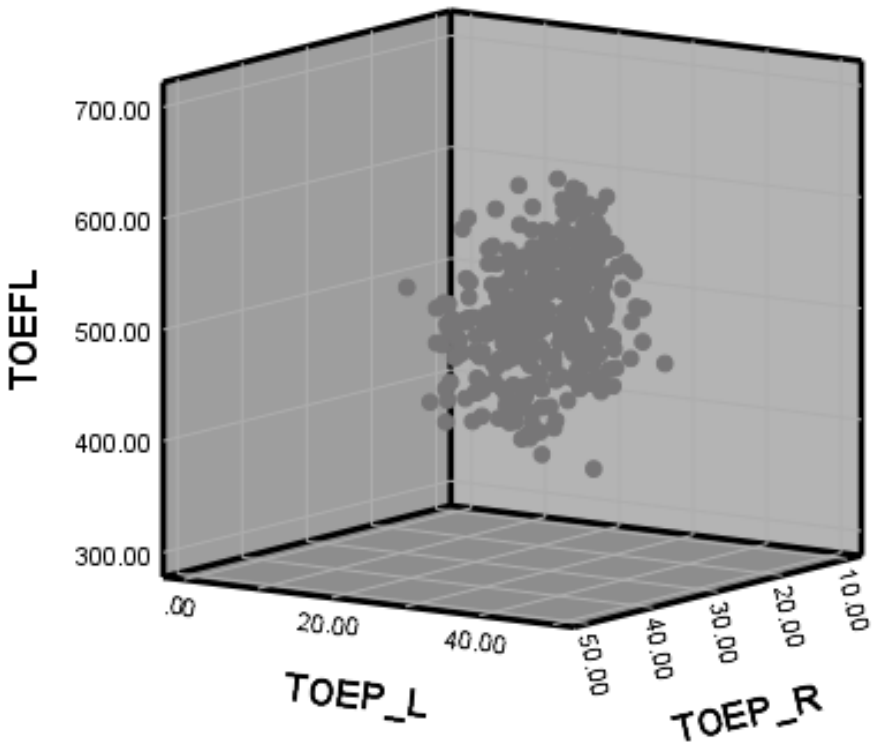


Table 1
The correlation, contribution and prediction equation of the TOELF scores with the TOEP scores as the predictor

Predictor	Dependent Variable	Cor-relation	Contri-bution	Prediction Equation	RMSE
TOEP_R	TOEFL_R	0.797	0.635	TOEFL_R=0.555*TOEP_R+33.613	
TOEP_L	TOEFL_L	0.829	0.688	TOEFL_L= 0.628*TOEP_L+34.367	
TOEP	TOEFL	0.892	0.796	TOEFL = 3.479*TOEP +293.076	29.669
TOEP_R and TOEP_L	TOEFL	0.895	0.801	TOEFL =2.761*TOEP_L+4.311*TOEP_R+288.712	29.302

RMSE: root mean square of error

The comparison of the prediction model by using the TOEP total scores and the combined scores of TOEP Listening and Reading indicated that the prediction model using the combined scores of TOEP Listening and Reading was a better model. This was indicated by the smaller RMSE (root mean square of error) of 29.302.

The results of analysis lead to a conclusion that TOEP is a strong predictor for the TOEFL scores. By using the best prediction, the TOEFL score can be calculated by the following formula: TOEFL Score TOEFL =2.761*TOEP_L+4.311*TOEP_R+ 288.712, and the TOEP variance contribution explains the TOEFL variance of 80.1% (a high category). For example, if a testee scores 35 in Listening and 40 in Reading, it can be predicted that the score that she/he will get when taking the TOEFL is 2.761*35+4.311*40+288.712 = 557.787.

By using the results of item analysis, the level of item difficulty was obtained. This level of difficulty was used to calculate the information function value and SEM. As an illustration, the results of the calculation for Listening and Reading of Form 81 are presented in Figure 5 and 6 respectively.

From Figure 5 it can be seen that the information function value of the ability increases until it reaches the maximum value, then drop again. By contrast, the standard error of measurement (SEM) decreases to reach the minimum value, then increases again. The two graphics (IFV and SEM) meet at the ability scale of -3.3 and +2.8. This means that the listening test is appropriate for test takers with the ability at the range of -3.4 to +2.9. In Figure 5 it can also be seen that the maksimum information function value lies at the ability scale of 0.0. This means that the the highest information function value is given to test takers with the ability scale of 0.0.

Figure 5
The information function value of listening

Copyright © 2020, *authors*, e-ISSN: 2502-6747, p-ISSN: 2301-9468

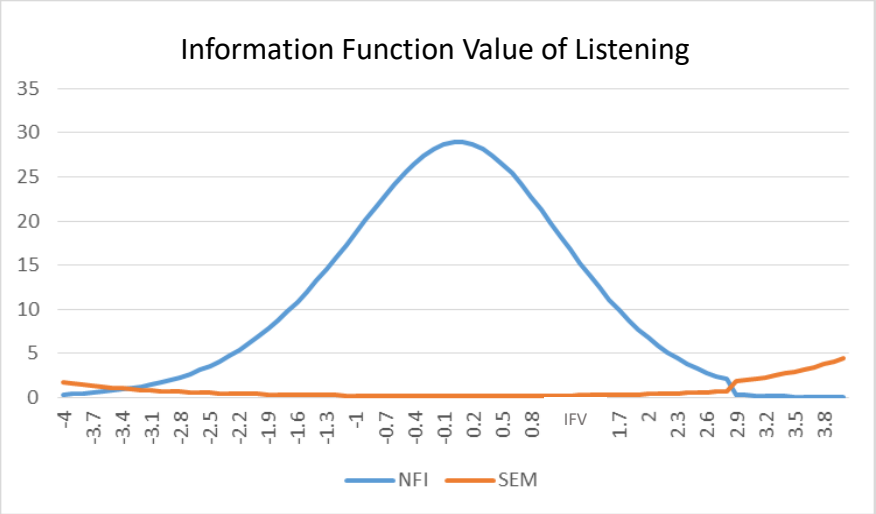
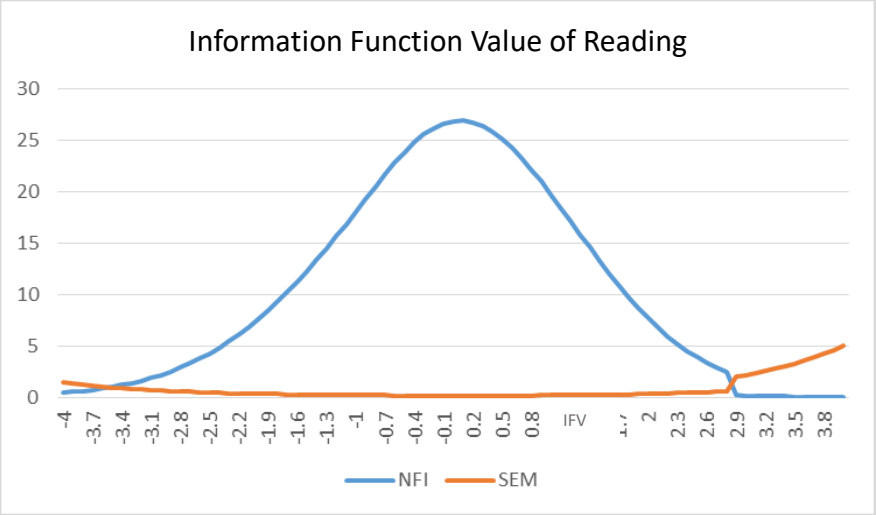


Figure 6
The information function value of reading



The relatively-same results were obtained from Reading, in which the information function value of ability increases to reach the maximum, then decreases, whereas the standard error of measurement decreases to reach the minimum value, then increases. These two graphics (IFV and SEM) meet at the ability range of -3.6 and +2.8. This means that the listening test used is appropriate for test takers with the ability range between -3.6 and +2.8. Figure 6 also indicates that the peak of information function value lies at the ability range of 0.0. This shows that the highest information function value goes with test takers with the ability scale of 0.0.

The range of information function values is bigger than SEM for each TOEP form and this is summarized in Tables 2 and 3. These tables show the lower value of -4.0 and the highest 4.0 in the z-score (-4+4) for both Listening and Reading.

Concerning the TOEP scores of both Listening and Reading, the scores to be obtained by test takers are minimum 0 and maximum 50. By using the conversion, the TOEP test takers' scores of Listening and Reading can be estimated by using the TOEL-L or TOEP_R formula = 6.25 (z-Score) + 25. The results of the conversion are presented in Table 3. This table indicates that the TOEP Listening and Reading Tests can measure well the test takers' ability with the lowest score of 2.5 and the highest score of 50.

Both the TOEP Listening and Reading scores can be converted into the TOEFL Listening and Reading scores by using the results presented in Table 4. Based on Table 4, the test takers' ability is well measured for the TOEFL Listening at the minimum of 34.160 and maximum of 69.692, whereas for Reading the minimum is 30.787 and maximum 58.625 as presented in Table 4.

By using the information on Table 4, the TOEFL scores can be predicted by using the TOEP Listening and Reading scores. The prediction results are presented in Table 5. This table shows that the prediction of the TOEFL scores measured well by TOEP range from the minimum of 310 and maximum of 656.34. In this study, the participants ability to obtain a TOEP Score was estimated using

Copyright © 2020, *authors*, e-ISSN: 2502-6747, p-ISSN: 2301-9468

IRT. The score was subsequently rescaled to 0-50. In Table 5, the TOEFL score can be predicted with TOEP. Since the Top limits of TOEFL scores are obtained from the test agency, we did not use IRT for TOEFL, but we used the final scores instead.

Table 2
The range of information function values for each TOEP form

Year	2016						2017				2019			
Set	77	77A	78	79	80	81	A	B	C	D	1	2	3	4
Listening														
Bottom Limit	-3.3	-3.3	-3.6	-3.2	-3.3	-3.3	-3.6	-3.6	-3.7	-3.5	-1.63	-0.60	-2.11	-1.48
Top Limit	2.8	2.8	2.8	2.8	2.8	2.8	2.9	2.9	2.9	2.9	2.53	2.22	4.33	4.00
Reading														
BottomLimit	-3.3	-3.3	-3.6	-3.2	-3.3	-3.3	-4	-3.5	-3.4	-3.6	-1.90	-1.62	-1.37	-4.0
Top Limit	2.8	2.8	2.8	2.8	2.8	2.8	-2.9	2.9	2.8	2.9	3.27	2.06	3.17	3.88

Table 3
Range of Test-takers' ability that can be measured by a good TOEP (Scale 0-50 for Listening and Reading)

Year	2016						2017				2019			
Set	77	77A	78	79	80	81	A	B	C	D	1	2	3	4
Listening														
Bottom Limit	4.375	4.375	2.5	5	4.375	4.375	2.5	2.5	1.875	3.125	14.813	21.25	11.813	15.75
Top limit	42.5	42.5	42.5	42.5	42.5	42.5	43.125	43.125	43.125	43.125	40.813	38.875	50	50
Reading														
Bottom Limit	4.375	8.75	5	10	8.75	8.75	0	3.125	3.75	2.5	13.125	14.875	16.4375	-1.25
Top Limit	42.5	42.5	42.5	42.5	42.5	42.5	43.125	43.125	42.5	43.125	45.438	37.875	44.813	49.25

Table 4
The Range of Test Takers' Ability Measured Well (in the TOEP score) for both Listening and Reading

Year	2016						2017				2019			
Set	77	77A	78	79	80	81	A	B	C	D	1	2	3	4
Listening														
Bottom Limit	35.72	35.72	34.55	36.11	35.72	35.72	34.55	34.55	34.16	34.94	43.67	47.71	41.79	44.26
Top limit	59.47	59.47	59.47	59.47	59.47	59.47	59.86	59.86	59.86	59.86	59.99	58.78	67.06	69.69
Reading														
Bottom Limit	33.65	36.52	34.06	37.34	36.52	36.52	30.79	32.83	33.24	32.43	40.90	41.87	42.74	32.92
Top Limit	58.63	58.63	58.63	58.63	58.63	58.63	59.03	59.03	58.63	59.03	58.83	54.63	58.48	60.95

Table 5
The Range of Test Takers' Ability Measured Well by TOEP Put in the TOEFL Score Range for Listening and Reading

Year	2016						2017				2019			
Set	77	77A	78	79	80	81	A	B	C	D	1	2	3	4
Bottom Limit	314.38	328.33	310.00	334.75	328.33	328.33	310.00	310.00	310.00	310.00	386.19	411.51	392.19	326.81
Top Limit	584.34	584.34	584.34	584.34	584.34	584.34	588.77	588.77	586.78	588.77	597.28	559.33	625.64	656.34

DISCUSSION

This study set out with the aim of finding out the range of scores measured through TOEP administered in 2016 and 2017. The most obvious finding to emerge from the current study was that the results of the correlation estimation showed that the three predictors (TOEFL_L, TOEFL_R, and TOEFL) were positively correlated to the four dependent variables (TOEP_R, TOEP_L, TOEP, and TOEP_R and TOEP_L). Of the four predictors, the TOEP predictor and the combined predictor of TOEP_L and TOEP_R showed the highest correlation compared to the TOEP_R, TOEP_L predictors, i.e. .895. In accordance with the present results, previous studies by Rethinasamy and Chuah (2011) demonstrated that the positive correlation between the predictor variables and the criterion variables denotes the accuracy of the predictor variables in predicting the criterion variables. It is also encouraging to compare this finding with Retnawati's (2016) finding that the higher the correlation between two varibales is, the more accurate the predictor variables will be in predicting the criterion variables. A possible explanation for these results may be that the four predictor variables possess a high level of accuracy in predicting the

Copyright © 2020, *authors*, e-ISSN: 2502-6747, p-ISSN: 2301-9468

four criterion cvariables, but the TOEP predictor and the combined TOEP_L and TOEP_R possess a higher level of accuracy in predicting the TOEFL score of the test takers.

Comparison of the findings with those of other studies by Rethinasamy and Chuah (2011) and Zheng and De Jong (2011) also confirms that a positive correlation between the predictor variables and the criterion variables is the evidence of the examination and moderation to ensure the test accuracy in measuring what to be measured. In other words, the high correlation coefficient (r= .895) provides evidence of the criterion validity of the concurrent validity type for TOEP and the combined TOEP_L and TOEP_R with TOEFL as the criterion. Interestingly, the correlation coefficient between the TOEP_L and TOEP_R predictor is not higher than the correlation between the TOEP_L and TOEP_R predictor. This indicates that the listening test at the two types of test possesses the common variance compared to the Reading test.

With respect to the results of the regression estimation, one interesting finding is that in explaining the variance of the criterion variable (TOEFL), the combined scores of TOEP_L and TOEP_R is a stronger predictor compared to the predictor using the TOEP total score, although the difference in contribution of both is not significant (.01%). However, if seen from the RMSEA value of both models, the smallest RMSEA value is obtained from the model using two predictors (TOEP_L and TOEPL_R), smaller than that obtained by the model with one predictor. These results suggest that the TOEP scores are a strong predictor to predict the TOEFL scores of the test takers, with the best prediction model using two predictors (TOEP_L and TOEPL_R). These results corroborate the findings of a great deal of the previous work by Retnawati (2016). It seems possible that the high correlation coefficient obtained from the prediction model using the predictors is the multidimensional content in the predictor variables, i.e. multidimensionality on the test kit that measures language competence, related to content, listening, reading, speaking, and writing. In this case only 2 are measured, Listening and Reading. This is well-grounded since TOEP is a proficient test of English consisting of listening and reading, of which both have different constructs.

What is surprising from this study is that the highest contribution of the predictor with the combined scores of TOEP_L and TOEP_R is worth 80.01% in explaining the variance of the TOEFL scores. This may indicate that about 8.89% of variance of the TOEFL scores cannot be explained by the predictor with combined scores of TOEP-L and TOEP_R. If the contribution of the predictor is partially examined, the contribution of TOEP_L and TOEP_R falls into the medium category (68.8% and 63.5%). This indicates that about more than 20% of TOEP_L and TOEP_L variance cannot be explained by the two aspects (TOEP_L and TOEP_R). Consistent with the literature, there are four likely reasons why the criterion variance (TOEFL) cannot be explained by the predictor, i.e. the effect of the item difficulty, the assessment method which in this case is related to differences in format or content types, speed to respond to test items, and test takers' background (differences in experience) (Wilson & Graves, 1999; Wilson et al., 2004). In other words, the low correlation coefficient or contribution of TOEP_L and TOEP_R to the TOEFL score could be attributed to the relatively easier items or relatively more difficult items and the scarcity of opportunity in using Engish in social interactions might have led to

Another important finding from this study is that based on the item analysis using the IRT-Rasch Model, the information function value is bigger than the standard error of measurement (SEM) when it is in the range of -3.7 (bottom limit) and 2.9 (top limit) in the z-score (-4+4), 2.5 (bottom limit and 43.125 (top limit) on the TOEP scoes, and 310 (bottom limit) and 656.34 (top limit) on the TOEFL scores. This means that the TOEFL scores can be predicted accurately by the the combine TOEP_L and TOEP_R when the TOEP_R and TOEP_L scores are in the range of 2.5 and 50 or the TOEFL score predicted is at the range of 310 and 656.34. If the combined scores of TOEP_L and TOEP is less than 2.5 and more than 43.125, the TOEFL score predicted will not be accurate because the TOEFL score predicted will be less than 310 and higher than 656.34 (in this range of scores the standard error of measurement will be bigger tha the information function value). These findings are in line with those of previous studies by Retnawati (2016) and Desjardins and Bulut (2017). These results are likely to be related to the facts that when SEM is bigger than the IFV, the information related to test takers' ability obtained through the measurement instrument will be inaacurate.

One unanticipated finding was that the inaccuracy of the TOEFL scores predicted by the TOEP scores not more than 656.34 may indicate that some TOEP items need improvement in terms of the quality. This stands to reason because the test information function is obtained from the sum of the item information function (Desjardins & Bulut, 2017). The item information function value is very much influenced by the quality of items. This finding is consistent with that of Wu et al. (2016) who found that items with a low discriminating power will lower the test reliablibity, increase the measurement error, and cause the test scores to be difficult to interpret or to be less meaningful. There are three likely possible reasons that might be related to the low discriminating power. One possible explanation is that the item measuring things other than the intended. Another possible

Copyright © 2020, *authors*, e-ISSN: 2502-6747, p-ISSN: 2301-9468

explanatiois is that the item is presented or constructed in a wrong way which in turns make the test takers confused. The last possible explanation is that the item has too high level of difficulty (too difficult) or too low level of difficulty (too easy) (Wu et al., 2016). To improve the accuracy of TOEP scores, it is therefore important to conduct an evaluation aimed at improving the quality of TOEP items.

CONCLUSIONS

This study provides a unique insight into the testing of criterion validity of the concurrent validity type of a measurement instrument and the information function value (IFV) and standard error of measurement (SEM) to find out the range of TOEP scores which can predict well the TOEFL scores of the test takers. It is unique because the problem of predictive validity with more standardized device criteria is rarely raised, due to the administration and the high expenses. Besides, this study is also important in relation to TOEP recognition. The discussion of the the results of the analysis leads to a conclusion that TOEP possesses the criterion validity against TOEFL, with the combined scores of TOEP_L and TOEP_R explaining relatively well the score variance of TOEFL of 80.1% and the TOEFL scores being predicted well are within the range of minimum 310 to maximum 656.34.

In general, the findings of this study indicate that the quality of TOEP items needs improving to ensure that the TOEP scores can perfectly explain the TOEFL scores, i.e. from the lowest score to the highest score of TOEFL. For further studies, the researcher can use the item response theory with the three parameter logistic (the 3-PL model) so that the information obtained is more extensive, or by comparing the minimum and maximum TOEFL scores which can be predicted by TOEP when the item parameter and ability of test takers is estimated using the item response theory with the 1-PL model, 2-PL model, and 3-PL model.

REFERENCES

Ahn, H. (2015). Assessing proficiency in the National English Ability Test (NEAT) in South Korea: A critique of a government's approach to testing English proficiency. *English Today, 31*(1), 34–42. https://doi.org/10.1017/s0266078414000522

Desjardins, C. D., & Bulut, O. (2017). *Handbook of educational measurement and psychometrics Using R*. CRC Press.

Fernandes, H. J. X. (1984). *Evaluation of educational programs.* National Education Planning, Evaluating and Curriculum Development.

Hambleton, R. K., & Swaminathan, H. (2013). *Item response theory: Principles and applications*. Springer Science & Business Media.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage Publications. Inc.

Hiser, E. A., & Ho, K. S. T. (2016). C-tests in Vietnam: An exploratory study of English proficiency. *Electronic Journal of Foreign Language Teaching*, *13*(2), 184-202.

Lee, J. S. (2006). Linguistic constructions of modernity: English mixing in Korean television commercials. *Language in Society*, *35*(1), 59–91. https://doi.org/10.1017/s0047404506060039

Madya, S., Retnawati, H., Purnawan, A., Putro, N. H. P. S., & Apino, E. (2019). The equivalence of TOEP forms. *TEFLIN Journal*, *30*(1), 88-104. http://dx.doi.org/10.15639/teflinjournal.v30i1/88-104

Mohammadi, M., & Enayati, B. (2018). The Effects of lexical chunks teaching on EFL intermediate learners' speaking fluency. *International Journal of Instruction*, *11*(3), 179–192. https://doi.org/10.12973/iji.2018.11313a

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Danish Institute for Educational Research.

Rethinasamy, S., & Chuah, K. M. (2011). The Malaysian university English test (MUET) and its use for placement purposes: A predictive validity study. *Electronic Journal of Foreign Language Teaching*, *8*(2), 234–245. https://doi.org/10.2139/ssrn.2146007

Retnawati, H. (2014). *Teori respons butir dan penerapannya*. Parama.

Retnawati, H. (2016). *Validitas, reliabilitas dan karakteristik butir*. Parama

Saeedi, M., Tavakoli, M., Rahimi Kazerooni, S., & Parvaresh, V. (2011). Do c-test and cloze procedure measure what they purport to be measuring?: A case of criterion-related validity. *International Scholarly and Scientific Research and Innovation 5*(2), 190–199.

Saito, Y. (2019). English language teaching and learning in Japan: History and prospect. In Kitamura, Y., Omomo, T. & Katsuno, M. (Eds.) *Education in Japan: A comprehensive analysis of education reforms and practices* (pp. 211–220). Springer

Sawaki, Y., & Sinharay, S. (2018). Do the TOEFL iBT® section scores provide value-added information to stakeholders? *Language Testing*, *35*(4), 529–556. https://doi.org/10.1177/0265532217716731

Simion, M. O. (2012). The importance of teaching English in the field of tourism in universities. *Annals-Economy Series, 2*, 152–154.

Copyright © 2020, *authors*, e-ISSN: 2502-6747, p-ISSN: 2301-9468

Wilson, K. M., & Graves, K. (1999). Validity of the secondary level English Proficiency Test at Temple University-Japan. *ETS Research Report Series*, *1999*(1), i-67. https://doi.org/10.1002/j.2333-8504.1999.tb01809.x

Wilson, K. M., Nagara, S. K. N., & Woodhead, R. (2004). TOEIC®/LPI relationships in academic and employment contexts in Thailand. *ETS Research Report Series*, *2004*(1), i-26. https://doi.org/10.1002/j.2333-8504.2004.tb01943.x

Wu, M., Tam, H. P., & Jen, T.-H. (2016). *Educational measurement for applied researchers. Theory into Practice*. Springer.

Zheng, Y., & De Jong, J. (2011). *Establishing construct and concurrent validity of pearson test of english academic*. Pearson Research Note. Retrieved from http://pearsonpte.com/research/research-summaries-notes/.

Copyright © 2020, *authors*, e-ISSN: 2502-6747, p-ISSN: 2301-9468