

## Item parameters of Yureka Education Center (YEC) English Proficiency Online Test (EPOT) instrument

<sup>1</sup>Endrati Jati Siwi; <sup>\*</sup>1Rosyita Anindyarini; <sup>1</sup>Sabiqun Nahar

<sup>1</sup>Yureka Education Center Yogyakarta

Jl. Palem Hijau No. 120, Sidoarum, Godean, Sleman, Yogyakarta 55264, Indonesia

<sup>\*</sup>Corresponding Author. E-mail: [anin@eurekatour.com](mailto:anin@eurekatour.com)

*Submitted: 1 April 2020 | Revised: 29 April 2020 | Accepted: 15 May 2020*

### Abstract

Yureka Education Center (YEC) is one of the institutions which has developed an online-based English proficiency test. The test is called the English Proficiency Online Test (EPOT) which follows the TOEFL ITP (Institutional Testing Program) framework. Thus, this study aimed to analyze the characteristics of EPOT instruments consisting of Listening, Structure, and Reading subtests, which later the quality of each EPOT test item is identified. This study used a descriptive quantitative approach by describing the characteristics of EPOT test items in terms of item difficulty index, item discrimination index, test information's function, and test measurement's errors. The data were collected through EPOT trials conducted by 2,652 online test-takers as participants from 20 provinces in Indonesia. The collected data were then analyzed using the Item Response Theory (IRT) approach using the BILOG program on all logistic parameter models which began with the item compatibility test against the model. Based on the results of the analysis, all subtests match the 3-PL model. Most of EPOT's test items had a good range of difficulty index and discrimination index. The EPOT information's function shows that accurate items are used on the 3-PL model for a certain capability range. This study is expected to point out that the EPOT test could be used as an alternative English proficiency test that is easy to use and useful.

**Keywords:** *analysis, parameter, EPOT, listening, structure, reading*

**How to cite:** Siwi, E., Anindyarini, R., & Nahar, S. (2020). Item parameters of Yureka Education Center (YEC) English Proficiency Online Test (EPOT) instrument. *REiD (Research and Evaluation in Education)*, 6(1), 51-65.  
doi:<https://doi.org/10.21831/reid.v6i1.31013>.



### Introduction

In this era of globalization or better known as free trade, each individual is required to prepare reliable skills, especially in the communication field. In the current situation, English has a big role related to global communication between countries. Therefore, each individual is expected to be able to master English actively both oral and written. As in Indonesia, English is one of the foreign languages learned at school. Nowadays, foreign languages, especially English, have an important role, especially in careers. The working world will give high appreciation to

the people who have good English ability (Handayani, 2016, p. 106). English ability is needed for various job positions, such as teachers, employees, receptionists, security guards, programmers, and job seekers. Many companies, government agencies, including the selection process for civil servant candidates (*Calon Pegawai Negeri Sipil* or CPNS) require English proficiency, one of which is proved by a Test of English as a Foreign Language (TOEFL) certificate (Arnani, 2019).

In addition to functioning as a requirement for studying abroad and applying for work, the usage of TOEFL in Indonesia has

an additional function as a test instrument. This addition gives a chance for several institutions to develop and organize a test measuring an individual's English proficiency level. Sharpe states that there are 180 countries that take the TOEFL test every year in language institutions spread throughout the world (Sharpe, 2002, p. 3).

Yureka Education Center (YEC) is one of the institutions which develop English proficiency tests as a test instrument following one of ETS products, TOEFL ITP (Institutional Testing Program). English Proficiency Online Test (EPOT) is a TOEFL Prediction Test which has been developed by YEC since 2018. As the name implies, EPOT measures an individual's English proficiency level in three aspects which are Listening, Structure and Written Expression, and Reading skills which can be done online.

EPOT gives several benefits for the test takers. One of the benefits is that the test can be done almost anywhere and anytime, as long as the test takers are connected to the internet. Moreover, the result of EPOT can be delivered instantly after the test ends. Test takers will receive a digital certificate sent to their registered email. EPOT is a web-based proficiency test, therefore, the test takers are not required to download any software or applications. They can take the test using a web browser on their laptops or personal computers.

EPOT has a test structure which refers to TOEFL ITP, consisting of three sections, namely: Listening Comprehension, Structure and Written Expression, and also Reading Comprehension. EPOT is held for 115 minutes. The exercises are in multiple-choice with

four answer choices. Table 1 is a comparison table of the number of questions and estimation time between TOEFL ITP and EPOT YEC.

To find out the quality of EPOT YEC test items, it is necessary to prove that each EPOT's test item is also capable of measuring someone's English proficiency as TOEFL ITP. The researchers verified each EPOT's test item using Item Response Theory (IRT) since the developed EPOT's test items do not depend on the ability of the test takers and vice versa. This means that the items' level of difficulty and discrimination do not depend on the test-takers (Anderson & Morgan, 2008, p. 76; Olufemi, 2013, p. 378; Yang & Kao, 2014, p. 171). In addition, Fan also said that the analysis using IRT emphasizes more on the level of test items' information, whereas, in classical test theory, the analysis emphasizes more on the level of the test's set information (Fan, 1998, p. 359). Thus, an analysis using IRT will give more detailed and accurate results (Pollard, Dixon, Dieppe, & Johnston, 2009, p. 3).

EPOT's items produce data with dichotomous scores in the form of correct (1) and incorrect (0). For dichotomous data, it can be analyzed using a latent linear model, perfect scale model, latent distance model, normal ogive parameter model, as well as the logistic parameter (de Ayala, 2009, p. 120; van der Linden & Hambleton, 1996, p. 18). This analysis of EPOT's test items chooses to use the parameter logistic model because the mathematical calculation is simpler using a logistic distribution model than using a normal distribution (Chung, 2005, p. 41).

Table 1. The Comparison between TOEFL ITP and EPOT YEC

Section	TOEFL ITP	EPOT YEC
Section 1: <i>Listening Comprehension</i>	Number of questions: 50 (35 minutes)	Number of questions: 50 (35 minutes)
Section 2: <i>Structure &amp; Written Expression</i>	Number of questions: 40 (25 minutes)	Number of questions: 40 (25 minutes)
Section 3: <i>Reading Comprehension</i>	Number of questions: 50 (55 minutes)	Number of questions: 50 (55 minutes)

Several previous studies about item analysis to measure the cognitive skills of the students used classical test theory. Still, the analysis using classical test theory did not yield enough information to find out the effectiveness of test items. The reason was the existing assumptions that could not be met. Item statistics depended on the test takers' characteristics and standard error of estimator score which applied to all of the test takers. Therefore, there was no estimator score for each of the test-takers and test items. Nowadays, there are several studies which are using IRT because this theory is considered to be more detailed and valid to reveal the test items' quality.

The main advantages of IRT are that (1) the item parameters are invariant function or the response curve unchanged; and (2) the item selection can be done based on the amount of item information and test information (Hambleton, Swaminathan, & Rogers, 1991, p. 7). According to Naga, there are two types of parameters that are related to one another. In this case, participant characteristic parameters can be known if the parameter characteristics of the items are known or also known as a logistic model estimation. This model estimation is then developed into a logistic model one-to-three parameter. Likewise, the parameter features of the items can be measured if the parameter characteristics of the participants are known as the maximum likelihood estimation or the estimation of the maximum probability of occurrence (Naga, 1992).

According to the logistic distribution, IRT model is classified based on the number of test item's parameter into three types namely one-parameter logistic model (1-PL), two parameters logistic model (2-PL), and also three-parameter logistic model (3-PL) (Hambleton, 1989, p. 148; Hambleton et al., 1991, p. 7; Magis, 2013, p. 305). The 1-PL model only has one parameter which is the level of difficulty; the 2-PL model has two parameters, namely, the level of item difficulty and discrimination index; while the 3-PL model displays the parameter of difficulty index, discrimination index, and also pseudo-guessing.

Item difficulty index (b) shows the difficulty level of an item. Item discrimination index (a) shows how each test item differentiates test takers' ability in answering that test item. Meanwhile, pseudo-guessing (c) shows the probability of test-takers with low ability to correctly answer a test item. In order to apply the theory, the researchers need to determine a suitable model with the analyzed data. For statistical model selection, from the three models, then the compatibility of the items was made based on the Chi-square values. If an item has a probability of the Chi-square value  $\geq 0.05$ , then that item is considered fit or compatible with the model. For this reason, the logistic model in data that has the most compatible items will be chosen as the model for data analysis (Retnawati, 2014, p. 25).

A research of the Test of English Proficiency (TOEP) developed by *Direktorat Pendidikan SMA* or the Directorate of Senior Secondary Education has been done by several researchers using Three-Parameter Logistics (3PL). It was in contrast with test items developed by private English courses. Currently, there are many institutions which offer online TOEFL Prediction test which can be easily accessed. However, the quality of test items they developed cannot be validated since it was not tested and evaluated properly. There were many test takers like college students or fresh graduates who have taken these tests to find out their English proficiency. As one of the institutions which develop TOEFL Prediction like test called English Proficiency Online Test (EPOT) and an online course, YEC makes serious efforts to analyze its test items using the IRT approach. This study was conducted to analyze and describe the parameter of EPOT's test items based on the parameter logistics which suited to the responses of EPOT's test-takers.

## Method

The study is aimed at finding out the parameters or the characteristics of EPOT's test items through the trial results. The parameter of EPOT's test items can be observed from the difficulty, discrimination, and also pseudo-guessing level of each test item. There

were 2,652 participants from 20 provinces throughout Indonesia which become the research subjects. Most of them are fresh graduates who wanted to apply for a job and students who wanted to continue their study. A simple random sampling technique was used in order to gather samples from the population. The samples were picked randomly neglecting any difference in the population. This method is used if the members of a population are considered homogeneous (Sugiyono, 2014). The samples were fresh graduate students from bachelor level with the minimum age of 23 years old. Most of the samples were taking EPOT since they needed a TOEFL certificate to apply for job vacancies or to continue their studies. Others were taking EPOT to test their proficiency level since EPOT's framework is equivalent to the TOEFL ITP.

All of the research subjects took EPOT online test through the official Yureka Education Center's website [yec.co.id](http://yec.co.id). A set of EPOT test consists of 50 listening comprehension questions, 40 questions of structure and written expression, and 50 questions of reading comprehension. The test should be done in 115 minutes. Previously, the testing of EPOT's validity and reliability has been conducted. The content validity testing was done by three English experts, examining the content and structure of the test. The results of the validity testing showed that there were four test items that were not valid since their Aiken's V index was less than 0.67 (Azwar, 2017, p. 113). These four items were then being revised and tested again to achieve a good Aiken's V index. The distribution of Aiken's V value is shown in Figure 1.

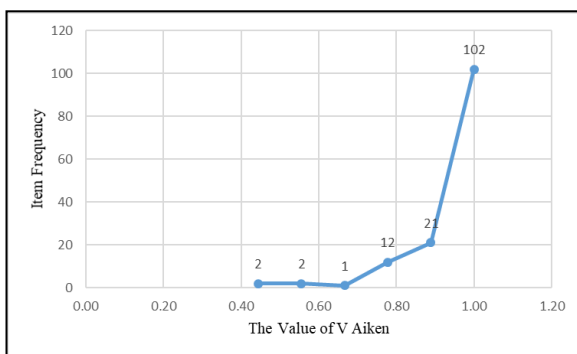


Figure 1. V Aiken Value Distribution

The face validity test was conducted by two experts on learning media. The experts examined the test appearance and the item context compatibility with the aim of the test. As the results, for the test appearance, YEC should add a button to change audio volume; recheck the audio playback; change the test instructions' placement; fix the test items' placement; fix the consistency of font size; and fix the writing whether it should be capital, italic, or bold. After the revision was done and the appearance of the test was improved, it can be considered that the face validity has been met (Azwar, 2017, p. 43). The reliability test of EPOT showed that it has Cronbach's Alpha score of 0.908. It meant that 90.8% of the observed score variant resembled the true score. According to the literature, the reliability score of 0.908 showed that EPOT's test instrument has good reliability (Gliem & Gliem, 2003; Guilford, 1956). Therefore, the developed EPOT's test instrument is assumed to highly reliable. The results of the reliability test are shown in Table 2.

Table 2. Reliability Index

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.908	.910	140

The item analysis on EPOT used the logistic parameter model. In IRT theory, the item's difficulty level can be labeled as good if the value is in the range -2 up to 2 (de Ayala, 2009, p. 15; Fan, 1998; Hambleton et al., 1991, p. 13). Theoretically, the item discrimination index is in the scale  $-\infty \leq a \leq \infty$ , but practically, the a value is in the range 0 up to 2 (Hambleton et al., 1991, p. 15). Meanwhile, c value was considered as a good item if it is in the range of 0 up to 1 or  $1/k$  that k is the total answer choices (Hulin, Drasgow, & Parsons, 1983). After going through the comparison process from the three logistic parameters, the 3-PL model was considered to be the most suitable model for EPOT trial result data.

The item analysis used Bilog-MG software. The computer program for maximum likelihood estimation was the Bilog-MG fit program that was used for one, two, or three-parameter model. The Bilog-MG program

was able to estimate multiple-choice items and also for estimating latent skills in huge amounts (Crocker & Algina, 1986, p. 354; Hambleton et al., 1991, pp. 43–50; Yen & Fitzpatrick, 2006, pp. 131–132). Based on the output of the Bilog-MG program, it could be obtained item difficulty index ( $b$ ) or threshold, item discrimination index ( $a$ ) or slope, and pseudo guessing ( $c$ ) or asymptote. The difficulty index, discrimination index, and the ability of items to be guessed by a participant will be shown in a graph. Besides, the Item Characteristics Curve (ICC) graph would show the quality of several items, and the Test Information Curve (TIC) graph will show the quality of EPOT.

### Findings and Discussion

EPOT consists of three sections, namely Listening Comprehension, Structure and Written Expression, and Reading Comprehension. The summary of difficulty index, discrimination index, and matched item can be seen in Table 3.

If the data are accumulated in 1-PL, there will be only 71 items from Listening, Structure, and Reading which has Chi-square  $\geq 0.05$ . In the 2-PL model, there are 117 items which have Chi-square  $\geq 0.05$ . Meanwhile, in the 3-PL model, there are 123 items which have Chi-square  $\geq 0.05$  or can also be considered as fit items. In conclusion, the logistic model that fits the EPOT test-takers answers results is the 3-PL model. The selection of the 3-PL model is also caused by some test-takers who already fulfilled the requirements for the use of the 3-PL model. Other than that, it also reinforces the assumption that proficiency tests using multiple-choice formats are examples of situations where the 3-PL model is suitable. Test takers tend to choose the best answer which they found most interesting if they could not find the correct answer, so the guessing factor is considered in this study (Huriaty, 2019, pp. 35–36).

Table 3. Summary of Item Parameters' Characteristics and Matched Item Analysis

Section	Model	Item's Description	Number of Good Item/ Item Fit	Percentage
Listening	1PL	$b$	49	98%
		Fit Item	27	54%
	2PL	$a$	45	90%
		$b$	48	96%
		Fit Item	45	90%
	3PL	$b$	46	92%
		$a$	50	100%
		$c$	10	20%
		Fit Item	48	96%
Structure	1PL	$b$	34	85%
		Fit Item	25	62.5%
	2PL	$a$	35	87.5%
		$b$	40	100%
		Fit Item	26	65%
	3PL	$b$	40	100%
		$a$	39	97.5%
		$c$	12	30%
		Fit Item	27	67.5%
Reading	1PL	$b$	44	88%
		Fit Item	19	38%
	2PL	$a$	46	92%
		$b$	45	90%
		Fit Item	46	92%
	3PL	$b$	44	88%
		$a$	49	98%
		$c$	3	6%
		Fit Item	48	96%

The first section, Listening, consists of 50 questions with a duration of 35 minutes. Based on the test-takers' response data, it is found out that EPOT Listening has various

difficulty index, discrimination index, and pseudo-guessing which can be seen in Figure 2, Figure 3, and Figure 4.

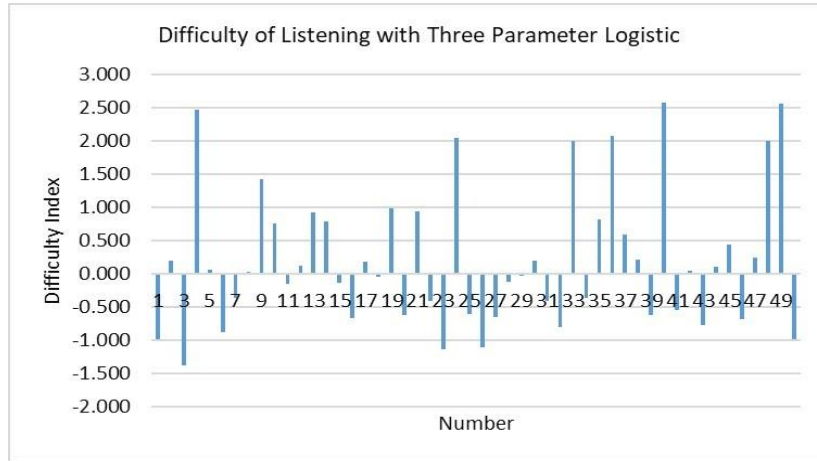


Figure 2. Difficulty Index of EPOT Listening

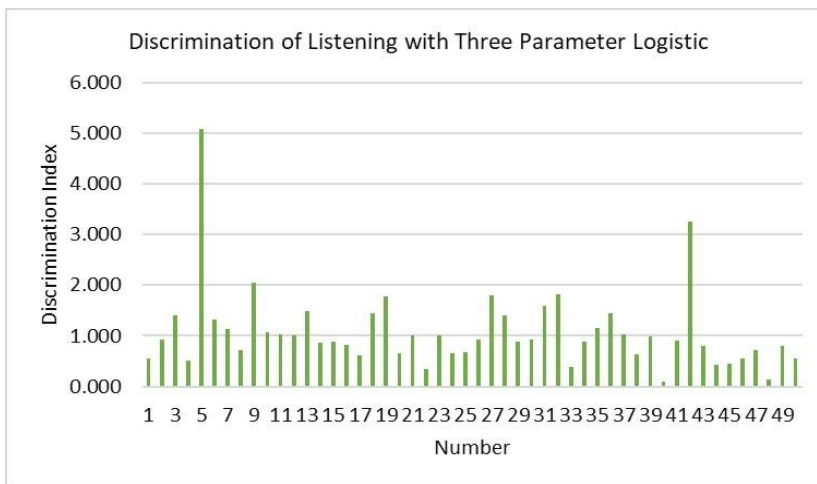


Figure 3. Discrimination Index of EPOT Listening

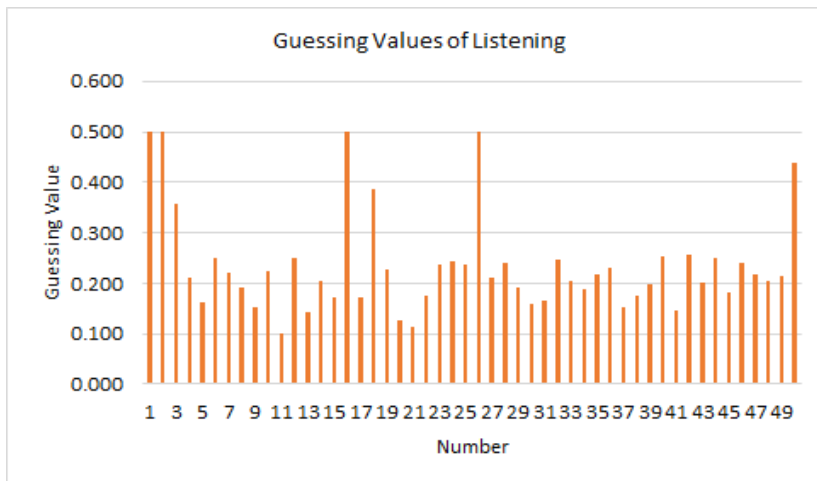


Figure 4. Pseudo Guessing Values of EPOT Listening

According to Figure 2, it can be concluded that there are 46 items out of 50 which have good difficulty index while four items are considered as poor. Those four items are number 4 ( $b = -2.473$ ), number 36 ( $b = 2.068$ ), number 40 ( $b = 2.572$ ) and number 49 ( $b = 2.552$ ). Number 36, 40 and 49 are considered too difficult because the  $b > 2$ , while number 3 is considered too easy because  $b < 2$ . It causes the answer responses' patterns tend to be poor and not able to show the difficulty index parameter. In Figure 3, it can be seen that the items in the Listening section have shown the various difficulty index and

are distributed well. All 50 test items show a good discrimination index with the range between 0 up to 2. Accordingly, the high and low ability of the test takers can be shown by the EPOT Listening test items.

On the other hand, Figure 4 shows that the Listening section has 43 items with good pseudo guessing. It means there are only 14% out of all items that can be answered correctly because there is an element of guessing. The next analysis is about the item fit analysis on Listening which gives an illustration in the form of Item Characteristic Curve (ICC) as presented in Figure 5 and Figure 6.

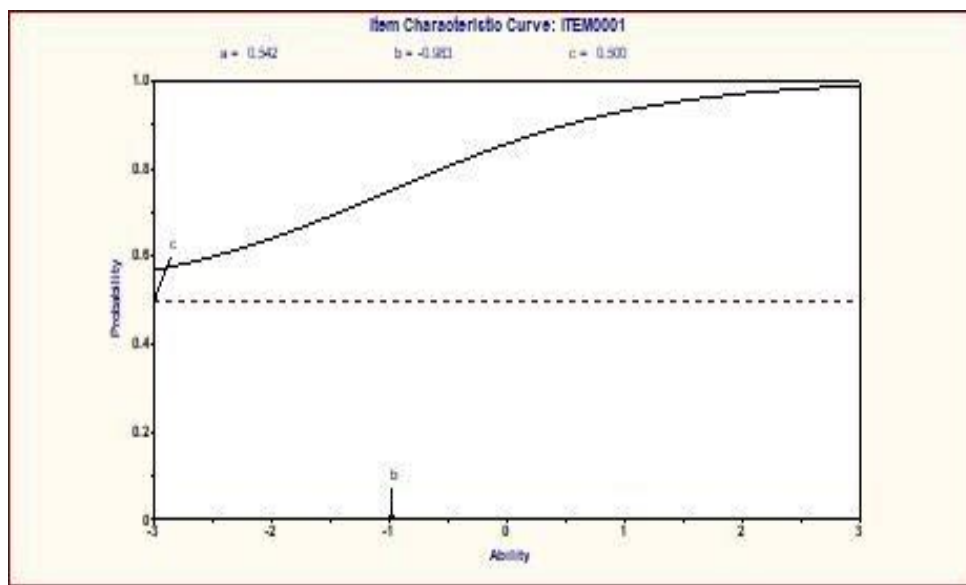


Figure 5. An ICC Example of Listening Item Number 1

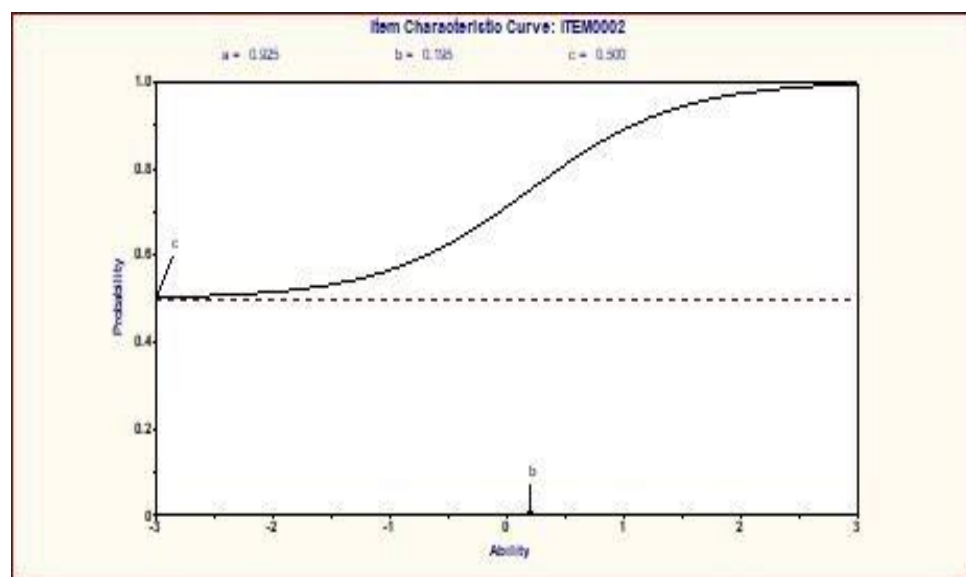


Figure 6. An ICC Example of Listening Item Number 2



Figure 5 and Figure 6 are examples of test-takers' responses pattern toward EPOT Listening test items number 1 and 2. Figure 5 shows a graph of the relationship between test takers' ability and parameter estimation item number 1 with  $b = -0.983$ ;  $a = -0.542$ ; and  $c = 0.500$ . Figure 6 illustrates the relationship between test takers' ability and parameter

estimation item 2 with  $b = 0.195$ ;  $a = -0.925$ ; and  $c = 0.500$ .

EPOT Structure section consists of 40 items done in 25 minutes. According to the data of test-takers' responses, 40 items of EPOT Structure also have various difficulty and discrimination index. These findings can be seen in Figure 7, Figure 8, and Figure 9.

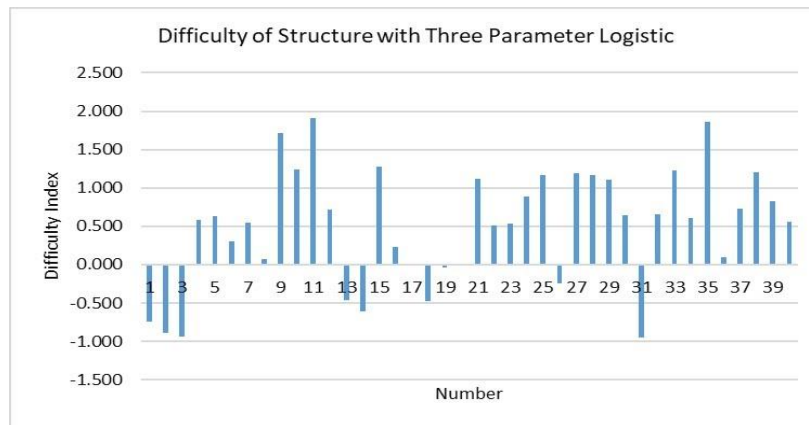


Figure 7. Difficulty Index of EPOT Structure

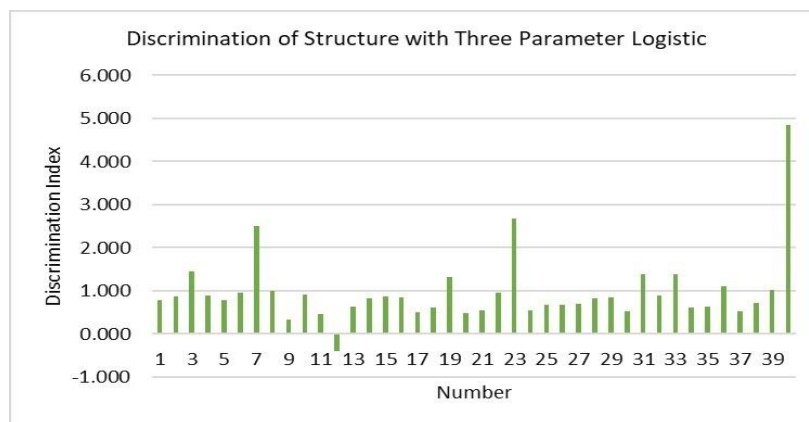


Figure 8. Discrimination Index of EPOT Structure

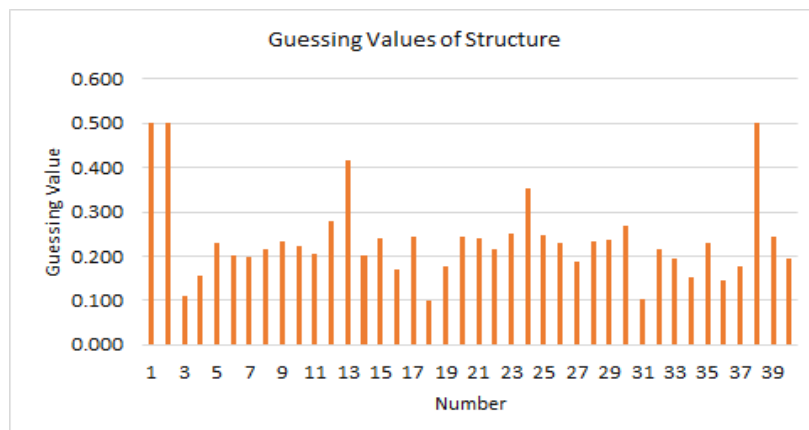


Figure 9. Pseudo Guessing Value of EPOT Structure



Figure 7 shows that all 40 EPOT Structure items have good difficulty level. In Figure 8, the 39 items have a good discrimination index. However, there is one item with a poor discrimination index, that is number 12 with a = -0.395. It shows that number 12 cannot show the difference between the low and high ability of the test takers. Meanwhile, Figure 9 shows that the Structure section has 35 items with good pseudo-guessing. In other words, there are only 12.5% out of all items that can be answered correctly because of the guessing element. The next analysis is about the item fit analysis on Structure, which gives an illustration in the form of ICC, as presented in Figure 10 and Figure 11.

Figure 10 shows the relationship graph of test takers ability and parameter estimation of item number 1 in Structure with  $b = 0.793$ ;  $a = -0.746$ ; and  $c = 0.500$ . Meanwhile, Figure 11 shows a relationship graph of test takers' ability and parameter estimation of EPOT Structure item number 2 with  $b = 0.879$ ;  $a = -0.893$ ; and  $c = 0.500$ .

The last section is Reading Comprehension. EPOT Reading section consists of 50 items that are done in 55 minutes. According to the test takers' responses, it can be concluded that 50 items of EPOT Reading also have various difficulty and discrimination index. It can be seen in Figure 12, Figure 13, and Figure 14.

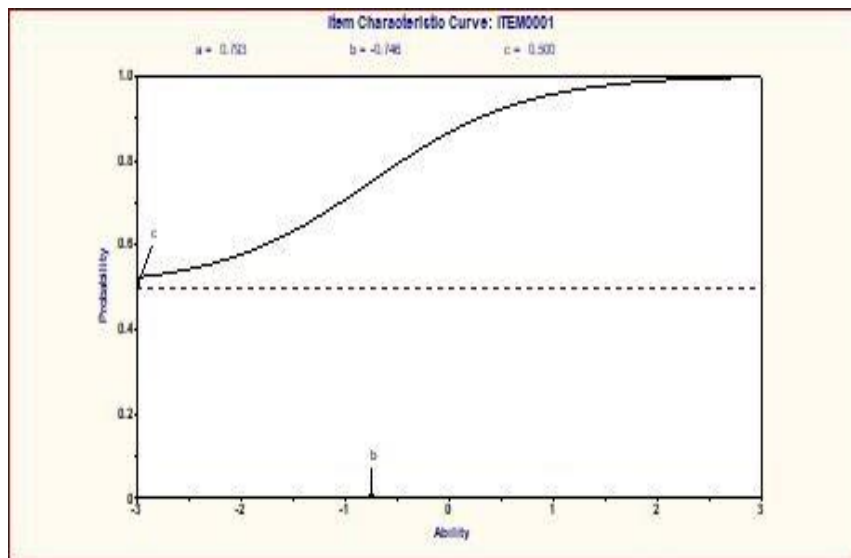


Figure 10. An ICC Example of Structure Item Number 1

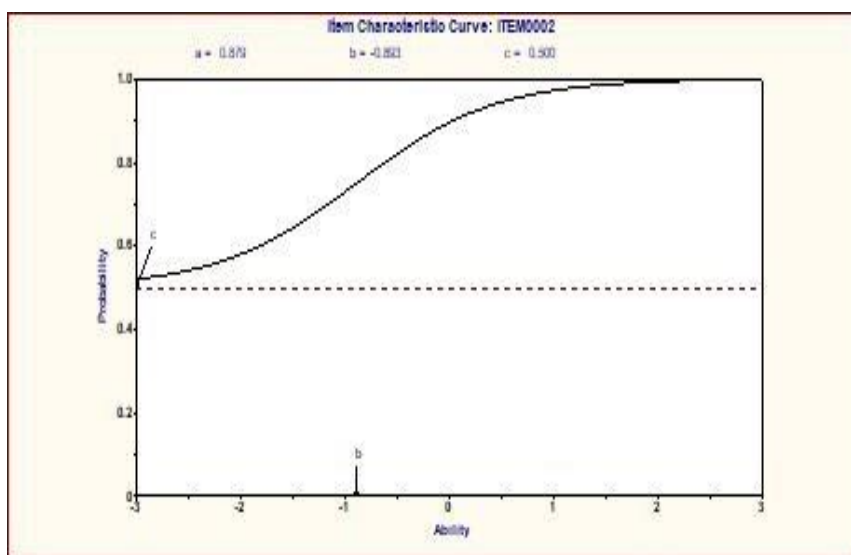


Figure 11. An ICC Example of Structure Item Number 2

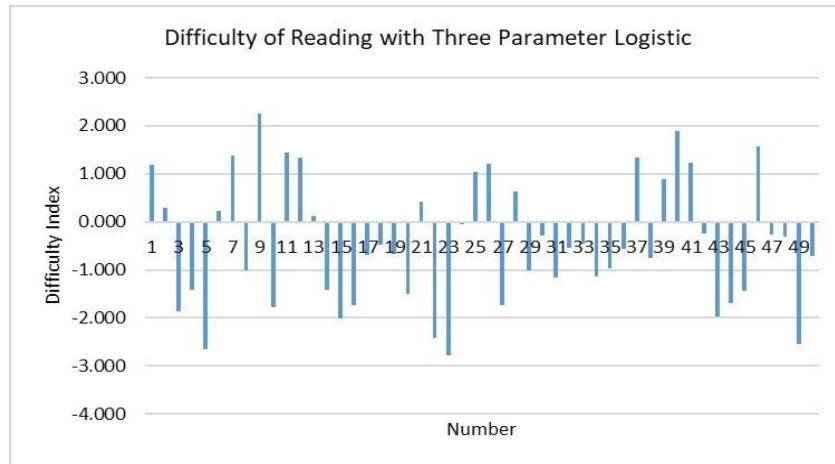


Figure 12. Difficulty Index of EPOT Reading

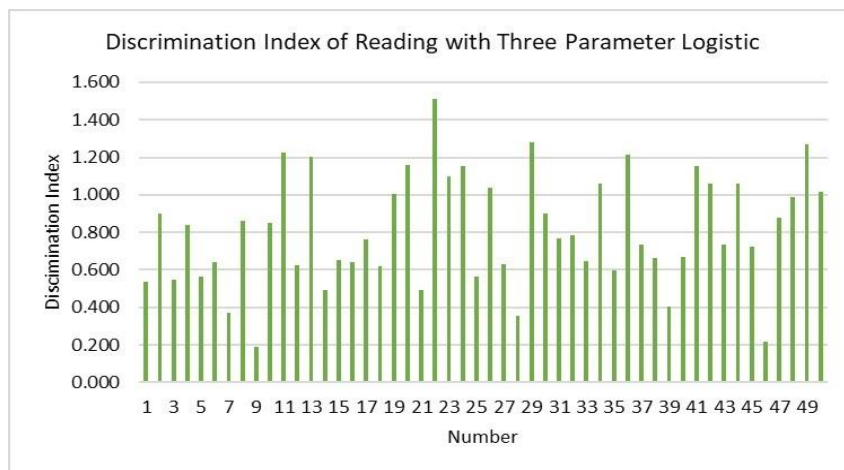


Figure 13. Discrimination Index of EPOT Reading

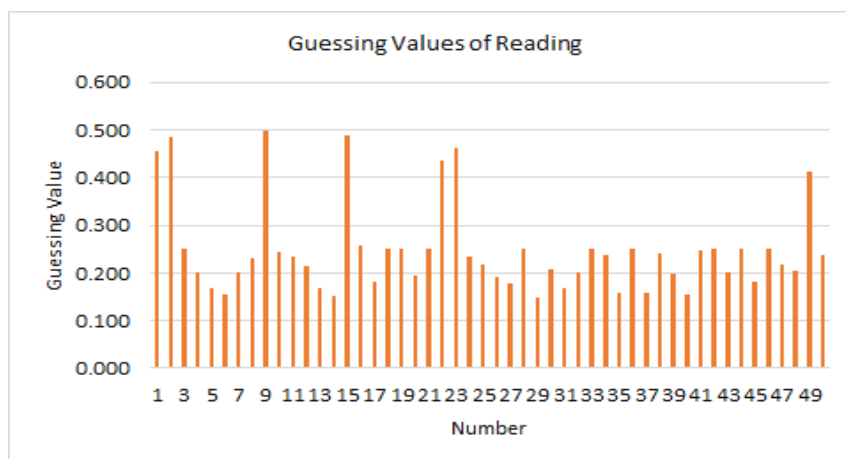


Figure 14. Pseudo Guessing Value of EPOT Reading

Based on Figure 12, 45 items have good difficulty index, and the remaining five items are considered poor. These five items are number 5 ( $b = -2.657$ ), number 9 ( $b = 2.264$ ), number 22 ( $b = -2.407$ ), number 23 ( $b = -2.771$ ), and number 49 ( $b = -2.547$ ). The

items number 5, 22, 23, and 29 are considered too difficult since the difficulty level is  $< -2$ ; and number 9 is considered too easy because the difficulty level is  $> 2$ . Thus, the test takers' responses tend to be poor, and these items cannot show the difficulty index parameter.

Figure 13 shows that all of the items in the EPOT Reading section have good discrimination index since they are in the range of 0 to 2 so that the test takers' low or high ability can be shown in all EPOT Reading's test items. Meanwhile, Figure 14 shows that the EPOT Reading section only has 43 items with good pseudo-guessing. It means 86% of all items can be answered correctly because of the guessing elements. The next analysis is about items fit in the EPOT Listening section, which gives an illustration in the form of ICC, as shown in Figure 15 and Figure 16.

Figure 15 shows a graph between the test takers' ability and estimated parameter Reading section item number 1 with  $b =$

0.536;  $a = 0.181$ ; and  $c = 0.455$ . In addition, Figure 16 depicts a graph between the test takers' ability and estimated parameter of EPOT Reading section item number 2 with  $b = 0.899$ ;  $a = 0.291$ ; and  $c = 0.484$ .

The next discussion will be about information function analysis and Standard Error Measurement (SEM). The EPOT information function value will show EPOT's reliability and measurement accuracy. The EPOT information function describes a low curve that increases, reaching the highest score in the middle before falling far from the midpoint. The curve's width shows the extent of the effective capability from the measurement results.

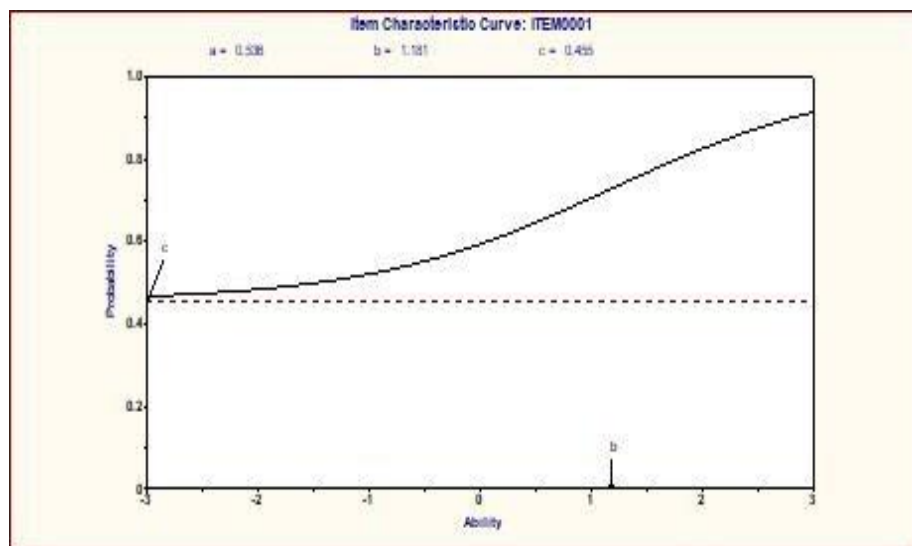


Figure 15. An ICC Sample of Reading Item Number 1

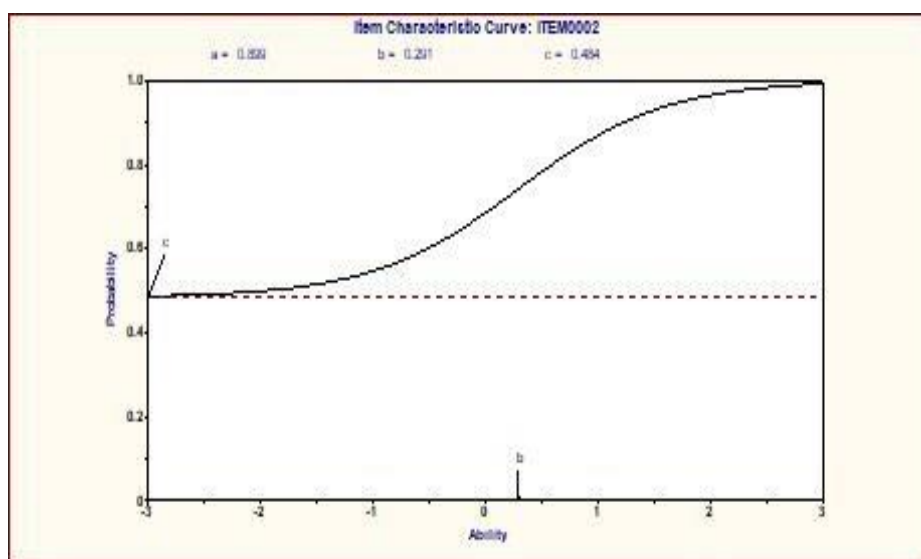


Figure 16. An ICC Sample of Reading Item Number 2

Test Information Function (TIF) will be effective if the curve line extends above the SEM line without having an intersection point. However, EPOT items' analysis yields

TIF and SEM curves that have interaction between the two. These are three figures which show the Total Information Curve (TIC) for 1-PL, 2-PL, and 3-PL model.

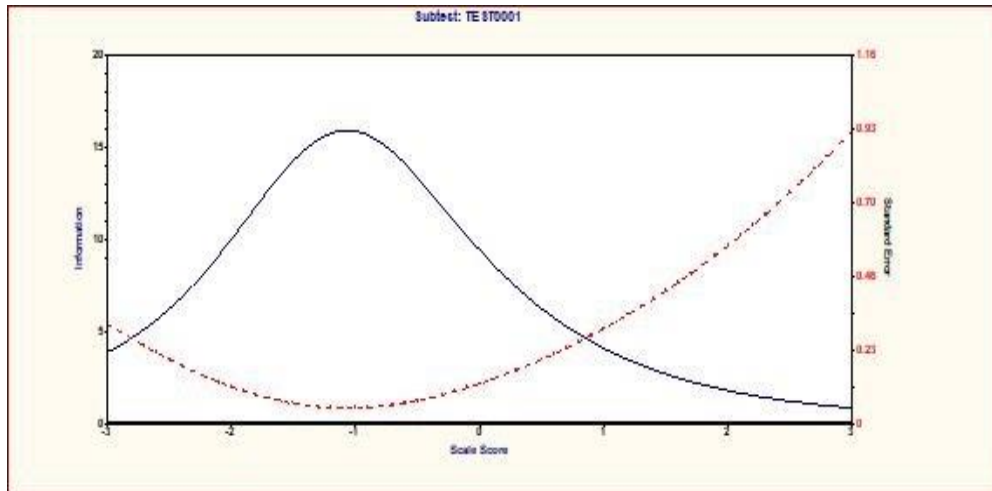


Figure 17. EPOT's TIC for 1-PL Model

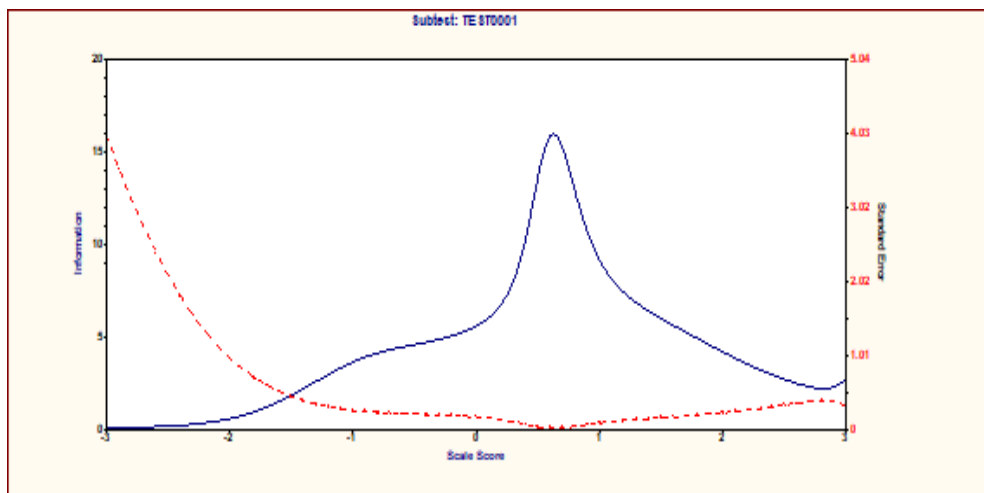


Figure 18. EPOT's TIC for 2-PL Model

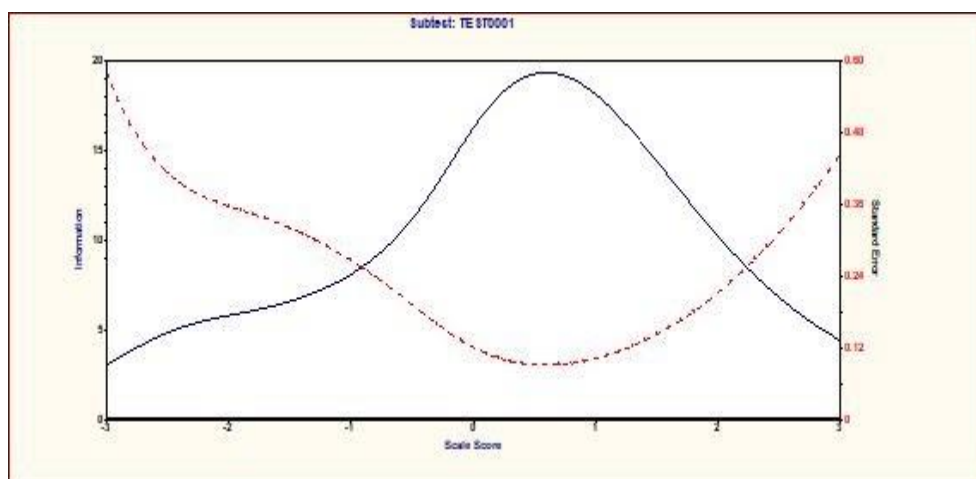


Figure 19. EPOT's TIC for 3-PL Model

Figure 17, Figure 18, and Figure 19 show TIC, which consists of the TIF line, SEM line, and interaction among them. TIC illustrates the total information produced by any level of ability. The dotted line shows SEM, which means the greater the information function, the smaller the measurement error is. The three graphs show the TIF curve above SEM with two intersection points; it means that the information obtained from the measurement results is only accurate on abilities with a certain range. This research's finding shows that the 3-PL IRT model provides the highest TIF compared to the 1-PL and 2-PL models. It is caused by the average of EPOT's items discrimination index with 3-PL model ( $a = 0.948$ ) higher than the item's discrimination index with 1-PL ( $a = 0.777$ ) and 2-PL ( $a = 0.460$ ). In the IRT model that accommodates the presence of discrimination index, if the discrimination index gets bigger, the value of TIF obtained will be greater (Setiawati, Izzaty, & Hidayat, 2018, p. 17; Yang & Kao, 2014, pp. 173–174; Zięba, 2013, p. 96). The presence of this discrimination index causes the item information with 2-PL is higher than 3-PL. As a result, the 1-PL model that becomes the lowest because this model does not accommodate the discrimination index parameter.

Based on the previous analysis, 93% of Listening, Structure, and Reading test item has a good average of difficulty index between -2 to 2. There are 10 test items that were considered poor; they were too difficult or too easy. These items were still used to vary the test items. As stated by Hingorjo and Jaleel (2012), test items with an average difficulty index are more desirable, test items with easy level can be placed in the beginning question as warming up, and the difficult item should be reviewed to avoid language confusion.

In addition, out of the 140 EPOT's test items, one item of Structure test and one item of the Reading test had a discrimination index of  $> 2$ . The two items are not modified since the gap between the scores and also the standard score is not significant. Meanwhile, the pseudo-guessing index showed that only 19 test items can be answered correctly by the test takers, which rely solely on guessing. The

results of TIF and SEM curved almost perfectly and interacted at two intersection points. The results of the study pointed out that the IRT 3-PL model provides higher test information function than the 1-PL and 2-PL model. The reason was the average of the EPOT's 3-PL discrimination index was higher than the 1-PL and 2-PL model.

## Conclusion

Item analysis can give useful information related to the item characteristics of a test set. English Proficiency Online Test (EPOT) is a set of English proficiency test developed by YEC and has gone through several processes of testing and evaluation on its test items. The testing and evaluation are using a 3-PL model to show the characteristics of the test, consisting of difficulty index, discrimination index, and pseudo-guessing index.

Based on the results of EPOT's item analysis using the IRT 3-PL model, it can be concluded that most of the items have a good difficulty index. Several items that have poor difficulty index are still used to vary the test items. Moreover, EPOT's test items are also able to effectively distinguish test takers' ability and improve test takers' reliability (Nelson, 2001; Wells & Wollack, 2003). Several test items that have poor discrimination index are not modified as the gap between the scores, and the standard score is not significant. As for the pseudo guessing index, there are only a few test items that can be answered correctly by the test takers who rely on guessing. In conclusion, EPOT has sufficient quality of effective test items, and it can be employed as a TOEFL Prediction test.

## References

- Anderson, P., & Morgan, G. (2008). *Developing tests and questionnaires for a national assessment of educational achievement* (V. Greaney & T. Kellaghan, Eds.). <https://doi.org/10.1596/978-0-8213-7497-9>
- Arnani, M. (2019, November 14). CPNS 2019, 9 instansi ini wajibkan TOEFL, berapa skornya? *Kompas.Com*. Retrieved from <https://www.kompas.com/tren/read/2019/11/14/120925265/cpns->



2019-9-instansi-ini-wajibkan-toefl-berapa-skornya?page=all

- Azwar, S. (2017). *Reliabilitas dan validitas* (4th ed.). Yogyakarta: Pustaka Pelajar.
- Chung, H. (2005). *Calibration and validation of the body self-image questionnaire using the Rasch analysis*. Master thesis, University of Georgia, Athens, GA.
- Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Harcourt Brace Jovanovich.
- de Ayala, R. J. (2009). *The theory and practice of Item Response Theory*. New York, NY: Guilford Press.
- Fan, X. (1998). Item Response Theory and Classical Test Theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, 58(3), 357–381. <https://doi.org/10.1177/0013164498058003001>
- Gliem, J. A., & Gliem, R. R. (2003). Calculating, interpreting, and reporting Cronbach's Alpha reliability coefficient for Likert-type scales. *Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education*, 82–88. Columbus, OH: The Ohio University.
- Guilford, J. P. (1956). *Fundamental statistics in psychology and education* (3rd ed.). New York, NY: McGraw-Hill.
- Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147–200). New York, NY: Macmillan.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.
- Handayani, S. (2016). Pentingnya kemampuan Bahasa Inggris dalam menyongsong ASEAN Community 2015. *Jurnal Profesi Pendidik*, 3(1), 102–106. Retrieved from [http://ispijateng.org/wp-content/uploads/2016/05/PENTINGNYA-KEMAMPUAN-BERBAHASA-INGGRIS-SEBAGAI-DALAM-MENYONGSONG-ASEAN-COMMUNITY-2015\\_Sri-Handayani.pdf](http://ispijateng.org/wp-content/uploads/2016/05/PENTINGNYA-KEMAMPUAN-BERBAHASA-INGGRIS-SEBAGAI-DALAM-MENYONGSONG-ASEAN-COMMUNITY-2015_Sri-Handayani.pdf)
- Hingorjo, M. R., & Jaleel, F. (2012). Analysis of one-best MCQs: The difficulty index, discrimination index and distractor efficiency. *JPMA: The Journal of the Pakistan Medical Association*, 62(2), 142–147. Retrieved from [https://jpma.org.pk/article-details/3255?article\\_id=3255](https://jpma.org.pk/article-details/3255?article_id=3255)
- Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. Homewood, IL: Dow Jones- Irwin.
- Huriaty, D. (2019). Analisis karakteristik parameter butir berdasarkan model Logistik 3 Parameter. *Lentera: Jurnal Pendidikan*, 14(2), 33–40. <https://doi.org/10.33654/jpl.v14i2.885>
- Magis, D. (2013). A note on the item information function of the four-parameter logistic model. *Applied Psychological Measurement*, 37(4), 304–315. <https://doi.org/10.1177/0146621613475471>
- Naga, D. S. (1992). *Pengantar teori sekor pada pengukuran pendidikan*. Jakarta: Gunadarma.
- Nelson, L. (2001). *Item analysis for test and surveys using Lertap 5*. Perth: Curtin University of Technology.
- Olufemi, A. S. (2013). Item Response Theory as a basis for measuring latent trait of interest. *Greener Journal of Social Sciences*, 3(7), 378–382. <https://doi.org/10.15580/GJSS.2013.7.062513691>
- Pollard, B., Dixon, D., Dieppe, P., & Johnston, M. (2009). Measuring the ICF components of impairment, activity limitation and participation restriction: An item analysis using classical test theory and item response theory. *Health and Quality of Life Outcomes*, 7, 1–20. <https://doi.org/10.1186/1477-7525-7-41>
- Retnawati, H. (2014). *Teori respons butir dan penerapannya: Untuk peneliti, praktisi*

*pengukuran dan pengujian, mahasiswa pascasarjana*. Yogyakarta: Nuha Medika.

- Setiawati, F. A., Izzaty, R. E., & Hidayat, V. (2018). Analisis respons butir pada tes bakat skolastik. *Jurnal Psikologi*, 17(1), 1–17. <https://doi.org/10.14710/jp.17.1.1-17>
- Sharpe, P. J. (2002). *How to prepare for the TOEFL test: Test of English as a foreign language* (10th ed.). Jakarta: Binarupa Aksara.
- Sugiyono, S. (2014). *Metode penelitian pendidikan: Pendekatan kuantitatif, kualitatif, dan R&D*. Bandung: Alfabeta.
- van der Linden, W. J., & Hambleton, R. K. (1996). *Handbook of modern item response theory*. [https://doi.org/10.1007/978-1-4757-2691-6 I](https://doi.org/10.1007/978-1-4757-2691-6_I)
- Wells, C. S., & Wollack, J. A. (2003). *An instructor's guide to understanding test reliability*. Madison, WI: University of Wisconsin.
- Yang, F. M., & Kao, S. T. (2014). Item response theory for measurement validity. *Shanghai Archives of Psychiatry*, 26(3), 171–177. <https://doi.org/10.3969/j.issn.1002-0829.2014.03>
- Yen, W., & Fitzpatrick, A. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). Westport, CT: American Council on Education and Praeger.
- Zięba, A. (2013). The item information function in one and two-parameter logistic models – A comparison and use in the analysis of the results of school tests. *Didactics of Mathematics*, 10(14), 87–96. <https://doi.org/10.15611/dm.2013.10.08>