Item Characteristics on Pro-TEFL Listening Section

Lina[1], D Mardapi[2], Haryanto[3]
[1,2,3]Yogyakarta State University
{[1]lina.2016@student.uny.ac.id, [2]djemari.@uny.ac.id, [3] haryanto@uny.ac.id}

Abstract. A good test should be proven theoretically and empirically. A test is considered as a good one if its items have appropriate content, language and construct based on theoretical analysis. Besides, the empirical analysis on the test items need to do in order to describe the quality of the test. Empirical item analysis can be done based on the framework of Classical Test Theory (CTT) and Item Response Theory (IRT). This study analyzed empirically the test item characteristics based on the 1 parameter logistic (1-PL) IRT framework. The items were analyzed based on the difficulty indices. Before analyzing the test items, assessment on IRT assumptions was done; unidimensionality and local independence. The data were gathered through testing. The responses from 334 test takers on 50 items of listening section on Pro-TEFL test administered by the Center for Language Development were used as empirical data. The assessment on IRT assumptions (unidimensionality and local independence) was done by using Factor Analysis which was assisted by SPSS for windows program. The item parameter (the difficulty index) was estimated by using Program R. Results showed that the assessment on IRT assumptions can be fulfilled. Then, the analysis of item characteristics was done to estimate the item parameter based on the 1-PL IRT model. Based on the analysis on the difficulty index, there were 2 items that were considered as poor items because the difficulty indices are more than +2 and they are considered as too difficult items.

Keywords: *Item Characteristics, Test, IRT, Item Parameter*

1 INTRODUCTION

The TOEFL is a standardized test to measure examinee' ability to use and understand English at the university level. The test is developed for non-native speakers who want to study in English language countries, especially in the United States. Since the test is designed for academic purposes, it evaluates how well examinees combine their reading, listening, speaking and writing skills to perform academic tasks [1]. A good test should be proven, both theoretically and empirically. A test is considered as a good one if its items have appropriate content, language and construct based on theoretical analysis. The empirical analysis on the test items also need to do in order to describe the quality of the test. Empirical item analysis can be done based on the framework of Classical Test Theory (CTT) and Item Response Theory (IRT) [2].

Item Response Theory (IRT) is a theory of measurement, more precisely a psychometric theory. It's a family of statistical model which can be used for demonstrating reliability and validity of measurement. IRT has been developed, during the last decades, as a new measurement system which become an important system to evaluate tests. IRT becomes an important complement of the preceded measurement theory. In analyzing the test items, CTT tends to test oriented rather than item oriented [3].

IRT describes the relationship between a latent trait, the properties of the items, and examinee's answers to the individual items. The examinee response to the test item is typically a mixture of his/her proficiency in the area that the test is covering and the difficulty of the particular item. Item Response Theory (IRT) is a method that attempts to enumerate these examinee and item characteristics in order to calculate the probability of the examinee in answering the item correctly [4].

ICEL 2019, March 23-24, Malang, Indonesia

DOI 10.4108/eai.11-7-2019.159630

The mathematical models employed in IRT specify that an examinee's probability of answering a given item correctly depends on the examinee's ability or abilities and the characteristics of the item. IRT models include a set of assumptions about the data to which the model is applied. Assumptions of the IRT model that should be hold by the data are unidimensionality, local independence, and parameter invariance [5]. The unidimensionality is a common assumption of IRT models that implies only one ability is measured by a set of items in a test. This assumption cannot be strictly met because there are several factors affecting test performances; e. g., motivation, test anxiety, tendency to guess the response. Factor analysis can be used to determine the dimensionality (i.e., number of factors) for the item responses in a test. If factor analysis identifies a single dimension (or factor), then the assumption of unidimensionality is met. The latent trait estimates are not test-dependent, and item parameters are not sample-dependent, but model-dependent.

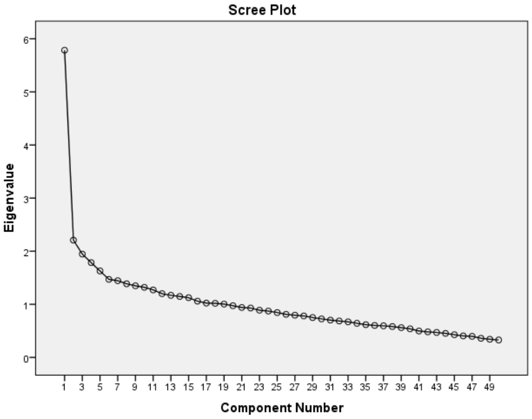
2 RESEARCH METHOD

This study analyzes empirically the characteristics of items on Pro-TEFL listening section of Center for Language Development UNY. The data were 334 examinee responses on 50 items which were gathered by testing. The items correct responses were denoted by ‘1’, while the wrong responses were denoted by ‘0’. Having the dichotomous data, the analysis were done through the following steps; (1) assessing the IRT assumptions, (2) determining the model analysis by assessing model-data fit, (3) estimating item parameter, and (4) analyzing the quality of the items based on the specified parameter. The first step is the assessment of IRT assumptions. The unidimensionality of the data were assessed by the factor analysis to find how many factor(s) measured by the test. [6] The factor analysis was done by using SPSS program for windows. The next step is assessing model-data fit to determine which model can explain the given data adequately. The assessment of model-data fit was done by calculating the Bayesian Information Criterion (BIC) which was assisted by program R. Having the appropriate model, the item parameter then was estimated based on the fit model. Estimating the item parameter was assisted by program R. The last step is analyzing the quality of the items based on the given parameter of the specified IRT model [1].

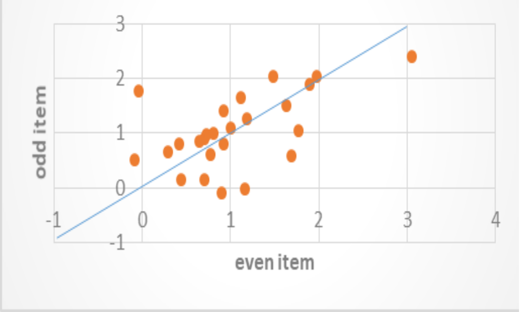
3 RESULT AND DISCUSSION

There are three assumptions that should be hold by IRT, namely uni-dimensionality, local independence, and parameter invariance. The results of the IRT assumption testing are described as follows [7]. The unidimensionality is the most widely used assumption to the IRT models. It implies that only one ability is measured by the items making up the test. Local independence implies that when the abilities influencing test performance are held constant, examinees' responses to any pair of items are statistically independent. In other words, after taking examinees' abilities into account, no relationship exists between examinees' responses to different items. These two assumptions can be tested by using factor analysis to determine how many factor(s)/dimension(s) measured by the test [8]. By analysing the output of factor analysis, which are in the form of eigenvalues and the screeplot, these assumptions can be proved.

A factor analysis was done to the 334 examinees responses of the 50 items on Pro-TEFL listening test. Results show that there were 19 test items or components which has the eigenvalue more than 1. Because of the great difference between the first component and the others; the eigenvalue of the first component is 5,8 while the others are less than 3, we can conclude that the pro-TEFL listening section test measures only one ability/dimension. Based on the dominant eigenvalue of the first component, the unidimensionality of the test can be proven. Below is the scree plot of the eigenvalues of 50 items/component resulted from the principal component analysis. The number of the peak shows the dimension or ability that are being measured.



The scree plot of the eigenvalues shows there is only one dominant factor measured by the Pro-TEFL listening test of Center for Language Development UNY. The dominant factor which is measured by the test is the examinees' listening skill. The test on parameter invariance can be done by analyzing the item parameter. The item parameter here are the difficulty index (b), the discriminant index (c), and the pseudo-guessing index(c). The researcher estimated the item parameter by using the program R. Each of the indices were broken into two parts, the indices of the odd and the indices of the even number. The indices of each parameter are plotted, and the line of X=Y are used to see the dispersion of the plotted indices. If the indices come near to the line of X=Y, so the item parameter can be considered as invariant. Below are the scatter plot of each parameter based on the odd and even number [9].



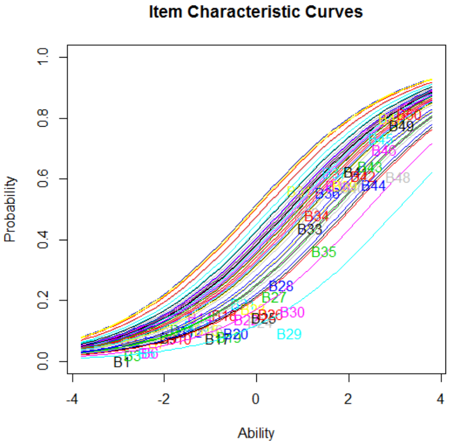
There are three mathematic models in IRT, based on the number of parameter specified. One parameter logistic model (1 PL) explains the data based on the item difficulty (b), two parameter model explains the data based on the item difficulty (b) and the discrimination index (a), while three parameter logistic model explains the data based on the item difficulty (b), the discrimination index (a) and the pseudoguessing index (c) [4].

To know which models fits to the data, we need to assess the goodness of fit firstly. If we use the unfit model, the given data can not be explained well by the model used. One of the way to assess the data goodness of fit is by calculating the Bayesian Information Criterion (BIC). The greater the value of BIC, the data less fit to the model, and vice versa. The value of BIC are calculated by Program R. The result of the program R analysis on the value of BIC of the three parameter IRT models (1-PL, 2-PL, and 3-PL) are presented in the following table.

Table.1 The Value of BIC of The Three Models

Model	BIC	Log.Lik
1-PL	20478.67	-10091.15
2-PL	20603.33	-10011.11
3-PL	20722.45	-9925.39

Table 1 presents the result of the calculation of BIC from the three IRT models. The lowest value of BIC and Log. Likelihood was obtained by the 1-PL, followed by the 2-PL and the 3-PL. As stated before, the greater the value of BIC, the data less fit to the model, and vice versa. Based on these results, we can conclude that the model which can adequately explain the given data is the 1-PL IRT model. This model becomes the basis in analyzing the item characteristics. The analysis of the item then be based on the item difficulty parameter (b).



The figure of the item characteristic curves shows the characteristic of the 50 items based on the difficulty parameter. The easier items are depicted in the left side, while the more difficult items are depicted in the right. Based on the ICCs of 50 items, we can see that item 29 is the most difficult item. The characteristics of the item were analyzed based on the model fit. An assessment of model-data fit found that the 1-PL IRT model is the most appropriate model for analyzing the given data. It means that the next analysis – item characteristic analysis of Pro-TEFL listening section, will be based on the item difficulty parameter. Program R was assisted to estimate the difficulty parameter (b). Result of the analysis showed that the difficulty indices of the 50 items ranging from -0.1014 to 3.0545. Theoretically, the range of difficulty parameter (b) of a good item is between -2 to +2. The greater the value of b parameter, the more difficult the item.

The difficulty index of item 29 (b29) is 3.0545, while the difficulty index of item 30 (b30) is 2.3986. Based on this rule, it was found that the two items (No. 29 and 30) are considered as poor items because these two are considered as too difficult items. Estimating and analyzing the item parameters, the Item Characteristic Curves (ICCs) then be made to describe the quality of the test items. The item characteristic curves (ICCs) are graphical depictions of the relationship between the measurement properties of the person and of the items parameter. The probability of examines in responding the item correctly ranges from 0.0 to 1.0 and the item difficulty (b) ranges from -4 to +4. The value of probability was scaled in y axis, while the ability (equivalence with item difficulty) was scaled in x axis.

4 CONCLUSION

A good test can help students improve learning and provide information exactly about their competencies. One of the criteria of a good test is that it must be able to differentiate the ability of each student. The higher the ability of students in understanding the subject matter, the higher the chance to answer the question correctly. The lower the ability of students in, the less chance to answer the item correctly. It is necessary to analyze the test empirically to provide data about the quality of the test.

This study found that the Pro-TEFL listening section developed by the Center for Language Development Yogyakarta State University contains good test items. Based on the IRT framework, the assessment of IRT assumptions was proven. The model fit assessment shows that the given data is adequately to be explained by the 1-PL IRT model in which the items were analyzed based on the difficulty parameter. Analyzing the items, it is found that there are 2 items which are considered as the poor item because they are too difficult to be answered correctly by the examinees.

REFERENCES

[1] R. K. Hambleton and H. Swaminathan, *Item response theory : principles and applications*. 1985.

[2] M. Djemari, *Pengukuran Penilaian & Evaluasi Pendidikan*. Yogyakarta: Nuha Media, 2012.

[3] H. Retnawati, “Proving content validity of self-regulated learning scale (The comparison of Aiken index and expanded Gregory index),” *Res. Eval. Educ.*, 2017.

[4] R. K. Hambleton and R. W. Jones, “Comparison of classical test theory and item response theory and their applications to test development,” *Educ. Meas. Issues Pract.*, 1993.

[5] A. Radford and T. Salimans, “Improving Language Understanding by Generative Pre-

Training," *arXiv*, 2018.

[6] R. Kuzar, "Constructions: A construction grammar approach to argument structure," *J. Pragmat.*, 2003.

[7] S. Sarjono, D. Mardapi, and M. Mundilarto, "Development of Physics Lab Assessment Instrument for Senior High School Level," *Int. J. Instr.*, vol. 11, no. 4, pp. 17–28, 2018.

[8] D. Manoppo, Yance; Mardapi, "An Analysis Of Method Of Cheating On Large Test Scale," *J. Penelit. Dan Eval. Pendidik.*, 2008.

[9] R. Hambleton and L. Patsula, "Adapting tests for use in multiple languages and cultures," *Soc. Indic. Res.*, 1998.