ITEM ANALYSIS OF ENGLISH SUMMATIVE TEST BASED ON ITEM RESPONSE THEORY (IRT)

*** Maria Lumbantoruan**
**** Indra Hartoyo, S.Pd., M.Hum.**

ABSTRACT

This study aims at finding out the indices of English summative test as well as describing the test quality based on the 3-PLM of IRT (Item Response Theory). A descriptive-quantitative design is applied by using Rstudio program to analyze the data. The findings show that there are three indices in evaluating English test based on the 3-PLM of IRT, they were item difficulty (*b* -parameter), item discrimination (*a* -parameter) and pseudo guessing (*c* -parameter). As much as 95 % (38 items) are in the good category of difficulty index and 5% (2 items) are in the poor category/very easy items. Then, as much as 30 % (12 items are in the good category and 70 % (28 items) in the poor category of the discrimination power. While referring to the guessing parameter, as much as 90 % (36 items) are in the good category and 10 % (4 items) are in the poor category.

Keywords: *Item Analysis, Summative Test, IRT*

INTRODUCTION

Assessment and evaluation are fundamentally needed in learning activity. They seem similar at a glance. However, basically, both of those terminologies do differ one another. According to Ekbatani & Pierson (2000); Lambert & Lines, (2000), assessment is a general item, which consists of methods and techniques that used to collect information about students' ability, knowledge and understanding as well as motivation. On the other hand, evaluation is the activity of gathering necessary information in order to determine the successful of the assessment in achieving its goal. It aims at discovering which methods work and which do not (Kaufman, Guerra, Platt, 2006).

In education system of Indonesia, assessment is given in form of test, either formative or summative test in each level of education. Final examination which belongs to summative test basically given to all students at the end of the semester. Commonly, after accomplishing this test, the teachers then make a decision whether or not the students are deserve to continue to the higher grade. In other words, in conducting an assessment, the test is expected to be able to assess the students' competence accurately. For that reason, in order to verify its quality, the test item should be verified by doing an evaluation.

According to Brown & Priyanvada (2010:9), evaluation is a process that leads to decision-making and solution- making for education process based on the result of tests, other assessments or others reports. In English language education, an evaluation is done in many facets of education including curriculum, teaching strategies, references and also test item.

In order to evaluate the test items, teacher need to do item analysis. Item analysis defined as a process that functioned to examine responses of the students toward test items (questions) and to assess the quality of those items and of the whole test. In conducting an items analysis, one of two approaches that can be applied is Item Response Theory (IRT).

Item Response Theory has several logistic models that used to estimate the item characteristics. Hambleton et al., (1991) and Baker (2001) describes that those models known as One-Parameter Logistic Model (1-PLM), Two Parameter Logistic Model (2-PLM), and Three-Parameter Logistic Model (3-PLM). The 1-PLM is an item response theory model which has only one parameter, exactly difficulty parameter (b). The items can be said to be good if it is neither too easy nor too difficult. The 2-PLM has two item parameters in the form of difficulty index (*b*) and discrimination index (*a*). Item that has high discrimination value will be able to discriminate high-ability examinees with low-ability examinees well. The 3-PLM has 3 parameters to be estimated, they are parameter of difficulty (*b*), discrimination power (*a*) and pseudo guessing parameter (*c*). The pseudo guessing parameter defined as the chances of low ability participants to answer a difficult item correctly by guessing. The good range of *c-parameter* is $0 \leq c \leq .35$ (Baker (2001: 37).

Unfortunately, even though the need of evaluation is very important, teachers usually are not aware of checking out the effectiveness of the test they made or given. As a matter of fact, based on an interview with the grade X English teacher of SMKS Parulian 1 Medan, the summative test items which had been administered to the students never been evaluated yet since the teacher only wants to obtain the score or the test result. In reference to the fact above, the researcher is interested to do an analysis in order to find out the indices of English summative test of SMK Parulian 1 Medan and describe its quality based on 3-PLM of IRT, as proposed by Baker (2001).

REVIEW OF RELATED LITERATURE

Assessment

According to Palomba and Banta (1999: 4), assessment means as the systematic collections, reviews, and uses of information concerning with educational program undertaken in improving both learning and development. Similarly, from the point of view of Sattler (1988), assessment is the relevant data collection which functioned for decisions-making. Furthermore, assessment refers to the systematic processes of measurement in terms of knowledge, behaviors, skills, attitudes, as well as beliefs based on explicit rules and benchmarks (Mahmoodi-Shahrebabaki, 2014, 2015).

Test

Brown (2000:384) stated that a method of measuring one's ability, knowledge or performance in a given domain is testing. It means that students' knowledge can be measure through testing. Additionally, the others stated that test is a tool or procedure used in measurement and assessment (Sudijono, 2007: 66; Arikunto, 2012:66). There are a number of test items. Usually test items are designed in many forms of questions. According to Day and Park (2005), forms of questions can be classify into 5, they are:

1. *Yes/no questions*

   Yes/no questions are simple form of questions that can simply be answered with either *yes or no* options. For example, *is this a famous book?* Generally, this form of test allows the student to possess a 50% chance of guessing the answer correctly.

2. *Alternative questions*

   Alternative questions are two or more *yes/no* questions connected with conjunction *or*, for instance: *Does the Bible tell only about the goodness of God or His holiness too?* Similar with *yes/no* questions, this question is prone to guessing.

3. *True or false*

The other form of questions is *True or False question.* Even though this kind of questions are commonly used, yet we cannot rely completely on them because it can be answer correctly by guessing too, without true understanding of the students.

3. *Wh- questions*

   Generally, every question that begins with *Wh-* such as *what, who, when, where, why and how* known as *Wh-Question*. This questions are very effective in measuring students' literal understanding of the text. It is very useful to identify their ability in recognizing important information from the text, conveying their personal responses or argument and predictions as well as in evaluation making. This test is also useful as a follow up questions for both alternative question and yes/no questions.

4. *Multiple-choice*

   Multiple-choice questions are designed with only a single correct answer added with some incorrect answers (distractors). In this multiple choice question, the form of the *Wh- questions* can also be used.

Evaluation

Fournier in Sibuea (2020) stated that evaluation can be done toward many things including in educational environment such as a program, product, person, policy, proposal or plan, curriculum, teaching strategies, references and test item. It is a process for gathering and drawing evidences that leads to a conclusion about the significance, value or merit, worth, and quality of one thing.

Item Analysis

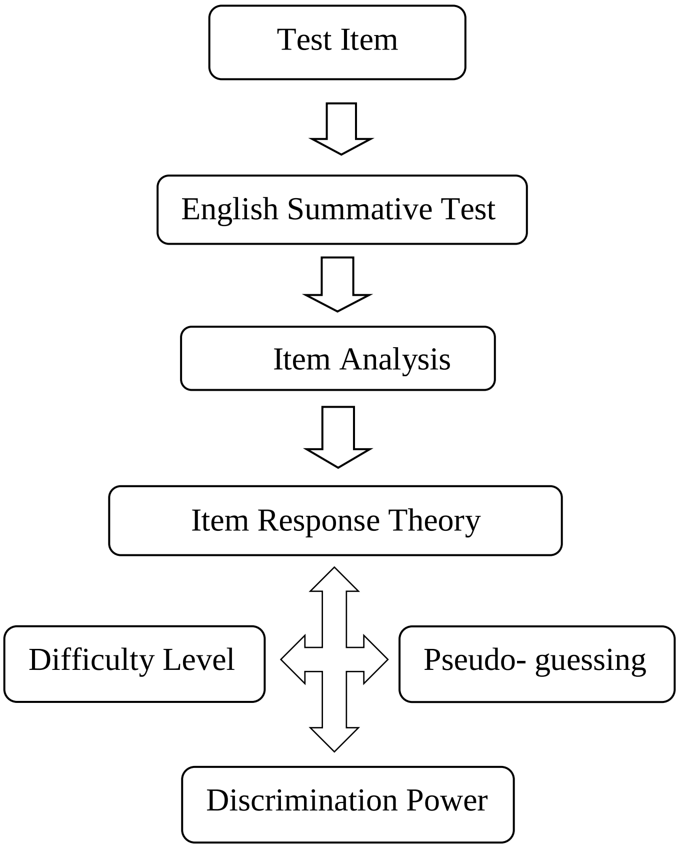
Item analysis is considered to be an important process in a test development that functions to provide information about item that need to be revised and even be maintained for the upcoming tests as well as to give information about items that should be discarded due to misleading (Quaigrain & Arhin, 2007). Briefly, some argued that item analysis is a set of procedures used to evaluate the test items quality (Musial, Nieminen, Thomas, & Burke, 2009).

Item Response Theory (IRT)

IRT that belongs to latent trait models provides a rich statistical tool to analyze educational test and psychological measurement scale. Hambleton & Jones (1993) described IRT as a general statistical theory concerning with test item and test performance as well as relationship among the performance with the abilities which are measured by the items in the test. Item parameters of IRT encompasses difficulty level, discrimination power and pseudo-guessing parameter. As stated by Hambleton et al., (1991) and Baker (2001), those parameters known as One-Parameter Logistic Model (1-PLM or Rasch model), Two-Parameter Logistic Model (2-PLM) and Three Parameter Logistics Model (3-PLM). 1-PLM encompasses only item difficulty level (b), the 2-PLM encompasses discrimination power (a) and the difficulty level (b) and the 3-PLM encompasses pseudo-guessing parameter (c) added with the discrimination power (a) and the difficulty level (b).

Conceptual Framework

In order to achieve a good test, evaluation toward the items should be done. One of ways to evaluate item characteristics or item quality is by doing item analysis. Since the English summative test of grade X students in SMKS Parulian 1 Medan has never been evaluated yet, thus in this research, it will be analyzed to find out its quality based on the 3-PLM of IRT. This research will be done through particular procedure according to application program that is used.



RESEARCH METHODOLOGY

This research is conducted by using descriptive design in which quantitative approach is applied. According to Gay in Syafitri et al (2017), the descriptive research is designed to describe the present condition of the research subjects. Further, quantitative method used to emphasize the analysis on numerical data (numbers) are processed with statistical methods (Azwar, 1999). In this study, the quality of test item is described by analyzing numerical data of test items quantitatively by using R studio program.

The data are information or facts used in discussing or deciding the answer of research question. The source of data in the study is the subjects from which the data can be collected for the purpose of research (Arikunto, 2010: 129). The data in this study are summative test items, which consist of 40 multiple choices items as well as students' response patterns in final semester taken from English teacher of SMKS Parulian 1 Medan.

FINDINGS AND DISCUSSION

Based on the data analysis, then it was found that There were three indices in evaluating English test based on the 3-PLM of IRT, they were item difficulty (*b* -parameter), item discrimination (*a* -parameter) and pseudo guessing (*c* -parameter). The difficulty index of the English summative test items in SMKS Parulian 1 Medan in 2020/2021 academic year ranges from -410 to -0.86, the discrimination index ranges from -0.00000000622 to +5.80 and the guessing parameter ranges from 0.00000000000000779 to 0.592. The distribution of the three parameters can be seen in the following table.

Table 4.1 Distribution of the Items Category Based On Difficulty Index

Difficulty Index	Category	Total Number	Item Number
b < -3	Poor (too easy)	2	34, 36
-3 ≥ b ≤ 3	Good	38	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24,

			25, 26, 27, 28, 29, 30, 31, 32, 33, 35, 37, 38, 39, 40
b > 3	Poor (too difficult)	0	-
Total Number		40	40

Based on the tables above, it was found that the quality of English summative test in for tenth grade students of SMKS Parulian 1 Medan in 2020/2021 academic year is as much as 95 % (38 items) were in the good category of difficulty index and 5% (2 items) were in the poor category/very easy items.

The items become poor because it constructed in a very basic level of Low Order Thinking Skills (LOTS). They refer to questions that measure the ability of students in terms of defining the meaning of a vocabulary. In blooms taxonomy, it refers to *remember* level. Consequently, the item is classified as a poor item because it is too easy. Meanwhile, the good one is constructed based on moderate level of LOTS, it is *understand* level.

Table 4.2 Distribution of the Items Category Based On Discrimination Index

Discrimination Index	**Category**	**Total Number**	**Item Number**
a ≤ 0	Unacceptable/ None	28	2, 3, 4, 5, 7, 8, 9, 11, 12, 14, 16, 17, 19, 20, 21, 23, 24, 25, 27, 28, 30, 31, 34, 35, 36, 37, 38, 39, 40

.01 - .34	Very Low	0	-
.35 - .64	Low	0	-
.65 – 1.34	Moderate	0	-
1.35 – 1.69	High	0	-
>1.70	Very High	12	1, 6, 8, 10, 13, 15, 18, 22, 26, 29, 32, 33
+ Infinity	Perfect	0	
Total Number		40	40

Then, based on the result of the discrimination power, as much as 30 % (12 items were in the good category and 70 % (28 items) in the poor category.

Table 4.3 Distribution of the Items Category Based On Pseudo-guessing Index

Pseudo-guessing Index	**Category**	**Total Number**	**Item Number**
0 ≤ c ≤ .35	Good	36	2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 14, 15, 16, 17, 18, 19, 20, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40
c > 3	Poor	4	1, 10, 13, 22
Total Number		40	40

While referring to the result of the guessing parameter, as much as 90 % (36 items) were in the good category and 10 % (4 items) were in the poor category.

CONCLUSIONS AND SUGGESTIONS

Based on the analysis result, then it can be concluded that in IRT, there are 3 parameter logistic models that encompass 3 item characteristics, namely difficulty level, discrimination power and pseudo-guessing parameter. Then, the result showed that some of the test items cannot be classified as good item since they cannot achieve the standard range of a good item as proposed by the theorist. Items with poor difficulty level occurred because they are constructed in a very low order thinking level.

For English teachers who usually construct test items/questions for students, it is suggested to be aware of the advantages of test item analysis to avoid the existence of such poor item. Then, it is so important to have good understanding in arranging good test items and to do evaluation through item analysis toward the test item that have already made and administered to the students in every testing. Then, for the next researchers, it is suggested to conduct such research with more explanation about the interrelationship among the 3 parameters logistics models and to do further action research that can become a new broader research, since there are still some weaknesses, lacks found in this research. It can be said that the result drawn has not been well-presented. It is due to researcher’s limited ability toward the topic. Therefore, this research is still so far from perfect.

REFERENCES

Arikunto, S. (2010). *Prosedur Penelitian Suatu Pendekatan Praktik*. Jakarta: Rineka Cipta.

Arikunto, S. (2012). *Prosedur Penelitian Suatu Pendekatan Praktek.* Jakarta: Rineka Cipta.

Azwar, S. (1999). *Metode Penelitian*. Yogyakarta: Pustaka Pelajar.

Baker, F. B. (2001). *The Basics of Item Response Theory (2nd Ed.).* College Park Maryland: ERIC Clearinghouse on Assessment and Evaluation. Retrieved from https://files.eric.ed.gov/fulltext/ED458219.pdf

Brown, H. D. (2000). *Principles of Language Learning and Teaching*. New York: Longman.

Brown, & Priyanvada. (2010). *Language Assessment: Principles and Classroom Practices*. New York: Pearson Education.

Day, Richard R. Jeong-suk Park. (2005). "*Developing Reading Comprehension Questions*". Reading in a Foreign Language, 17(1).

Ekbatani, G., & Pierson, H. (2000). *Learner-Directed Assessment in ESL Mahwah*. New Jersey: Lawrence Erlbaum Associates.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, California: Sage Publications.

Hambleton, R. K., & Jones, R. W. (1993). *An NCME Instructional Module on: Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development*. Educational Measurement: Issues and Practice, 12(3), 38-47.

Kaufman, R., Guerra, I., & Platt, W. A. (2006). *Practical Evaluation for Educators Finding What Works and What Doesn't.* California: Corwin Press. doi: 10.4135/9781412990189

Lambert, D., & Lines, D. (2000). *Understanding Assessment: Purposes, Perceptions, and Practices*. London: Routledge Falmer.

Mahmoodi-Shahrebabaki, M. (2014). *Using Self-Assessment Checklists to Make English Language Learners Self-Directed*. International Journal for Education, 3(6), 9-20.

Musial, D., Nieminen, G., Thomas, J., & Burke, K. (2009). *Foundations of Meaningful Educational Assessment*. New York: McGraw-Hill.

Palomba, C. A., & Banta, T. W. (1999). *Assessment Essentials*. San-Francisco, California: Jossey-Bass.

Quaigrain, K., & Arhin, A. K. (2017). *Using Reliability and Item Analysis to Evaluate a Teacher-Developed Test in Educational Measurement and Evaluation*. Cogent Education, 12(4), 1-11. doi: 10.1080/2331186X.2017.1301013

Sattler, J. M. (1988). *Assessment of Children*. San Diego: Jerome M. Sattler.

Sibuea, L.P.S. (2020). *Evaluating English Summative Test for Ninth Grade Students of SMPS Letjen S Parman Medan Based On Item Analysis*. Undergraduated Thesis, State University of Medan.

Sudijono, A. (2007). *Pengantar Evaluasi Pendidikan*. Jakarta: PT Raja Gravindo Persada.

Syafitri, R. et al. (2017). *The Students' Ability in Using Conjunctions (A Descriptive Quantitative Study of the Sixth Semester Students of English Study Program Bengkulu University*. Journal of English Education and Teaching (JEET), 1(1), 58-64.