CHAPTER I

INTRODUCTION

A. The Background of The Study

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). It means that evaluation refers to the way in which assessment verified, whether or not it is appropriate/good.

Banta & Palomba (2014) stated that assessment is a systematic process of providing credible evidence of resources, implementation, and learning outcomes. It is intended to improve and develop effective instructions, program, and services in higher level of education. Additionally, Sridadi (2007) also argues that assessment of student's learning outcomes intended as an attempt in which a variety of continuous and comprehensive information about the process and the outcomes of learning that have been achieved by the students through the learning activities are collected as the basis to determine the next steps. Briefly, it is evident that assessment has many purposes and contributions in learning activity.

1

Commonly, in every level of education, teachers or educators assess the students by giving them a test. According to Hughes (2003), a test is functioned to measure student's achievement and the degree of the success of the learning and teaching programs. Brown (2004) also stated that by giving a test, teachers eventually can measure one's ability, students' knowledge and performance in a particular domain. So, it constitutes a good move when teacher or even government design and implement such test for the students.

In education system of Indonesia, assessment is given in form of test, either formative or summative test in each level of education. This is regulated in Government Regulation (PP) No. 19, 2005 on National Education Standards, particularly article 63 (paragraph 1) which conveys that education both in primary and secondary level encompasses assessment of learning outcomes, educational units and government. Furthermore, in article 63 (paragraph 1) conveys that assessment is done on an ongoing basis which is functioned to monitor the process and progress as well as improvement of the results in the form of daily tests, midterm examination, and final examination in the end of the semester and also examination in order to continue to the next level or grade. So, it can be inferred that testing - as a way to assess students' ability - can be done through several forms.

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. The results of

the test should reflect the real condition of the students because the test results will be useful for decision-making.

Additionally, the percentage of the error (mistake) would affect the accuracy of learning outcomes of the assessment. Thus, in order to provide information accurately, assessments must fulfill a criteria. It should be valid, reliable, competence focused, comprehensive, objective, and educative (Puskur, 2008). On the other hand, the others argued that there are three concepts namely validity, reliability, and objectivity that must be verify to obtained the quality of the information from tests and other assessment instruments or methods (Anderson, 2003; Kubiszyn & Borich, 2013). This implies that the teachers must be aware of the accuracy and the quality of the test item in order to achieve the good criteria of the test.

Teachers or test makers cannot easily claim that the test they have already made is good enough or has a high quality before testing it to the student and obtaining the response of the student in answering each item of the test given then studying and analyzing the result of each item of the student's answer. 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. Furthermore, Madsen (1983:178) argues that a good evaluation of test is able to help teachers in measuring students' skills

in more accurate way. It also indicates that the teachers are aware of what they are teaching. Eventually, it can be inferred that an evaluation is an important thing to do for learning and teaching improvement.

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. According to Gronlund (1977), doing item analysis will contribute some advantages. First, it allocates data that helps the students to increase their learning. Second, it gives deeper insight and more skills for the teacher or the test maker to identify specific areas of course that need to be improved as a basis in preparing or constructing better test in the future testing. In other words, through item analysis, teachers can improve items that will be used again in later tests, and it can also be used to discard ambiguous and misleading items in a single test administration. Additionally, it will help teachers to determine whether or not the test appropriate for the students. In other words, by doing an item analysis, teachers can obtain information about the item suitability in relation with the average of the students' ability or competencies as a whole. However, this research does nothing about it specifically.

In conducting an items analysis, there are two approaches that can be applied. They are Classical Test Theory (CTT) and Item Response Theory (IRT). Classical Test Theory refers to a simple model that describes how measurement errors can influence observed scores. As a matter of fact, as stated by Hambleton et al in Aristiawan et al (2018), the most critical weakness of CTT is that the examinees' characteristics and characteristics of the test depend each other or

cannot be separated. In other words, the ability of the examinee is determined by the test. The examinees will seem to have low ability at the time they do difficult test item, while at the time the examinees face such easy item, they will seem to have higher ability. It means that the test characteristics are depend on the examinees or students' ability and vice versa.

On the other hand, IRT has some advantages, including: the item statistics is not dependent on the group, the obtained - test scores can illustrate individual capabilities and it can provide the right measurement for each ability score. According to Hambleton et al., (1991), Item Responses Theory builds on two basic postulates. First, the performance of examinees on test items can be estimated or explained by a series of factors known as trait, latent nature, or ability; the second is the relationship between the performance of the examinee and the set of characteristics of items can be explained by functions called Item Characteristics Curve (ICC). This curve describes that the higher the ability of the examinee, the higher is the probability of the respondent answering correctly to an item. Additionally, Naga in Fitriani et al., (2019) argues that the difficulty level of item and other characteristics will be invariant toward the examinee group when using IRT. It means that there is no problems with the subject who does the test. For that reason, Item Response Theory (IRT) then proposed as a solution to handle the weaknesses of Classical Test Theory (CTT).

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). Each

model has its own functions. It means that each of them has a different number of estimated parameters.

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. In ICC (Item Characteristic Curve), discrimination power is indicated by the slope of the curve. Steep slope of the curve refers to an item with high discrimination index and the flat curve plots the low discrimination item. 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. This probably happen because of lack understanding about the importance of evaluation or they do not recognize or know the way of evaluating tests or they feel that item analysis is too time-consuming (Mehrens, 1991: 160). As stated also by Danuwijaya (2018), item characteristics, such as item difficulty, item distractors, and other facets are not widely revealed by the test developers or language centers in Indonesia. Actually, it cannot deny that most of teachers or educators in Indonesia are rarely aware of the need of item test analysis itself.

They more likely focus on the test result than to do an evaluation toward the test given.

SMK Parulian 1 Medan, as one of vocational schools, conducted English summative test in 2020/2021 academic year. 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. Actually, by doing item analysis then test developer or teacher can eventually determine whether or not an item needs to be revised, to be maintained or even to be omitted or discarded. However, the teacher is rarely aware of the advantages of item analysis, and it means that the test quality has never been investigated yet.

Thus, in reference to the reason above, the researcher is interested to verify the quality of English test item that administer in the final semester examination through a research entitled: **"Item Analysis of English Summative Test Based on Item Response Theory (IRT)".**

B. The Problems of the Study

Based on the data given above, the problems of this study can be formulated as follows:

1. What are the indices of English summative test based on 3-PLM of IRT in SMK Parulian 1 Medan?

2. How is the quality of English summative test based on 3-PLM of IRT in SMK Parulian 1 Medan?

C. The Objectives of the Study

In reference to the problems above, this study aims at:

1. Finding out the indices of English summative test based on 3-PLM of IRT in SMK Parulian 1 Medan

2. Describing the quality of English summative test based on 3-PLM of IRT in SMK Parulian 1 Medan

D. The Scope of the Study

There are two approaches in evaluating the characteristics that reflect the quality of test item. This study focuses on the Item Response Theory in which the difficulty level and the discrimination power as well as the pseudo-guessing parameter become the main concern of the analysis. There are several types of assessment used in the learning activity for each lesson field as well as for each level of class. This study focuses only on English summative test that made by the teacher encompasses final semester test of 2020/2021 academic year. In administering an assessment, the teacher can provide or make some kinds of test item. This study focuses on multiple-choices test items.

E. The Significances of the Study

This research is expected to give theoretical and practical significances. Theoretically, this study can provide useful information for those who are interested in conducting such item analysis research in terms of item difficulty index, discrimination index and pseudo-guessing parameter. While practically, the researcher expects that this study will be useful for the English teacher as a reflection of the learning quality. It is expected to be an evaluation for the better construction of test. For the further researchers, this study is expected to be a relevant study for doing research especially concerning with English test evaluation that still very rare.