Indonesian EFL Journal: Journal of ELT, Linguistics, and Literature
||Volume||5||Issue||1||Pages|| 24 – 38 ||2019||
Available online at:http://ejournal.kopertais4.or.id/mataraman/index.php/efi |P-ISSN: 2460-0938; E-ISSN: 2460-2604|

DETERMINING THE QUALITY OF ENGLISH TEACHER-MADE TEST: HOW EXCELLENT IS EXCELLENT?

Nurhalimah
nurhalimah.ali14@mhs.uinjkt.ac.id

Syarif Hidayatullah State Islamic University of Jakarta, Indonesia
Jl. Ir. H. Juanda No 95 Ciputat 15412, Indonesia

Fahriany
fahriany@uinjkt.ac.id

Syarif Hidayatullah State Islamic University of Jakarta, Indonesia
Jl. Ir. H. Juanda No 95 Ciputat 15412, Indonesia

Dadan
dadan@uinjkt.ac.id

Syarif Hidayatullah State Islamic University of Jakarta, Indonesia
Jl. Ir. H. Juanda No 95 Ciputat 15412, Indonesia

Article History:
Received:
August 27, 2018

Revised:
March 25, 2019

Accepted:
July 10, 2019

Corresponding Author:
+81296662479

Abstract: Many studies have been conducted on evaluating the quality of a teacher-made test. Item analysis is crucial for making a good test, and improving test items. In response to the advantages of item analysis, this study looks at the theoretical and practical benefits of item analysis. The objectives were to know and to describe the extent of the quality of the English test items concerning difficulty level and discriminating power. This research used descriptive quantitative analysis. A total of 171 respondents of second-year students at MAN 1 Kota Tangerang Selatan 2017/2018 academic year were included in this study. The findings indicate that the English mid-term test has 24 acceptable items (80%) from the quality excellent, good, and satisfactory. Then, three items (10%) have poor quality, and three items (10%) have very poor quality, or in the negative value on discrimination index to the extent that the items are eliminated. It is proven by statistical data that they fail to distinguish between students who are knowledgeable and those students who are not on the base of how well they know the materials that have been tested.

Keywords – Teacher-Made Test, Item Analysis, Difficulty Level, Discriminating Power

INTRODUCTION

The most primary concern in the educational system probably is whether students achieve the goal of the education curriculum. In the Act of the Republic of Indonesia Number 20 on National Education, the curriculum is a set of plan and regulations on the purposes, content and material of lessons, and the method. It is as the guidelines for the implementation of learning activities to attain given education objective (Act of the Republic of Indonesia Number 20 on National Education, 2003). The education objective is referred to what students are exactly expected to be able to do at the end of a period of the process of teaching and learning. Evaluation is one crucial aspect that is closely related to the curriculum for determining the success or failure of teaching and learning process in the educational system. Evaluation has an essential role in the educational system since evaluation is conducted to make sure whether the overall teaching and learning processes have been running well throughout the process of teaching and learning.

Evaluation is paramount important in teaching and learning process. Evaluation is a systematic application of scientific methods to assess the design, implementation, advancement or results of a program (Desheng, 2013). The objective of the evaluation is to find out the extent to which learning has taken place. To evaluate teaching and learning process, the test is administered by teachers to their students as a part of the evaluation. The test is administered because teachers want to find out whether their students have mastered the content and material of lessons that have been taught in the teaching and learning process.

Concerning testing as a part of the evaluation, a test is often made by teachers themselves, or it is known as a teacher-made test. Particularly in language testing, there are specific qualities expected of a good language test which include validity, reliability, objectivity, and economy (Foyewa, 2015). To know whether a teacher-made test has fulfilled characteristics of a good language test, a teacher can evaluate the quality of a test on each test item after it has been administered to representative samples of their students. In doing so, item analysis is helpful for improving the quality of test items. In this case, item analysis is conducted through empirical judgment to ensure the quality of test items. The characteristics determined through this item analysis are item difficulty, item discrimination, and item distractor.

Item difficulty is the percentage of the test takers that marked an item correctly (Boopathiraj, 2013). In this case, the difficulty is referred to the relative frequency with which students taking the test chose a correct answer. Analyzing item discrimination is intended for distinguishing between students who are knowledgeable and those who are not from how well the students know the materials have been tested. Finally, item distractor determines how effective each alternative option is on multiple choice items.

Concerning the quality of test items with a teacher-made test, Quaigrain also questioned to what extent teacher-made test is reliable and valid. He proposes using reliability and item analysis to evaluate teacher-developed test in educational measurement and evaluation (Quaigrain, 2017). Furthermore, he suggests item analysis is crucial in improving test items which will be reused in later tests. Later, they can be kept in item banks. Also, Quaigrain (2017) states that item analysis can also be used to eliminate misleading items in a test. In response to this matter, Boopathiraj (2013) conducted an analysis of item difficulty and discriminating index on test items in the subject of Research in Education. Based on the findings, it showed that some items fail to distinguish between postgraduate students who are knowledgeable and students who are not knowledgeable about Research in Education subject. In his research, the test items are eliminated because the discrimination index falls in the poor category. Moreover, to develop the quality of test items, analysis on the item difficulty and discriminating power can be repeated in any subjects (Boopathiraj, 2013).

Criteria of a Good Language Test

In language testing, there are certain qualities expected of a good language test. The characteristics of a good language test include validity, reliability, objectivity, and economy (Foyewa, 2015).

a. Validity

A good test measures what it is supposed to measure. Validity is a crucial consideration in evaluating tests, and it is the most critical dimension of test development. Validity is the degree to which scores can be interpreted as a meaningful indicator of the construct of interest (Young, 2013). There are two basics categories of validity which include logical and empirical validity. Logical validity deals with logical judgment to ensure the validity of a test. On the other hand, empirical validity emphasized factor analysis based on correlations between test scores and criterion measures.

b. Reliability

According to the California Department of Education (2004), in language testing, an indicator of the extent to which scores are consistent across different administrations and/or different scores is defined as reliability. In the field of testing, it is not the test that is reliable, but the test score. If the test is administered to two groups of students with equal ability under the same testing condition, the results of the two tests should be the same, or very similar. Also, reliability is a general term used to describe measurement error. In this case, an error is defined as the differences in scores from the same test that has been given to the same students many times. This condition assumes that a student takes the same test and forgets each testing occurrence many times.

c. Objectivity

Foyewa (2015) argues that objectivity refers to the quality of a language test which ensures that a test should have one and only one correct answer. When the scorer does not need judgment in scoring, then the test is objective. The examples of objectivity are items in the form of "multiple-choice" and "true and false" test. In psychometric-structuralist movement, the objective testing in which the reliability (consistency of the score), validity (the representativeness of the sample) and objectivity (of test format) is preferred because they become the main concern (Sujana, 2000). Standardized test such as TOEFL is carefully constructed in objective formats so that they are easy to administer and score to meet objectivity of the test.

d. Economy

This quality of a language test ensures that the cost of administering a test, the time involved in setting and marking it should be equal or similar in degree with the expected outcomes obtained from it (Foyewa, 2015). In this quality of a good test, a test is not considered as economical if it takes much time, much energy, and cost much to construct.

Understanding Item Analysis

After a test has been administered to students and scored, one of the teachers' tasks is to evaluate the test's effectiveness that has been given to the students. This procedure often involves an analysis of each item on the test. In determining the effectiveness of individual items on the test items, item analysis is conducted. According to

Surapranata (2009), item analysis is generally conducted through two ways: qualitative control (*logical validity*) and quantitative control (*empirical validity*) on the purpose of finding out the usefulness of test items. Logical validity deals with analyzing the materials, construction, and other technical aspects based on logical thinking, meanwhile empirical validity deals with analyzing the items after they have been administered to representative samples to determine the effectiveness of the items based on empirical judgment supported by adequate statistical data.

Kinds of Item Analysis

Qualitative Analysis (Logical Validity)

Logical validity is determined based on logical thinking. A test is considered having logical validity when it is proven that after conducting analysis; the test logically measures what it is supposed to measure (Sudaryono, 2012). Logical validity emphasizes the quality of a test based on logical judgment. When a test logically measures what it is supposed to measure, then it can be said that the test has the criteria of logical validity. In determining whether a test item has logical validity, it can be done through analyzing two aspects; that are, content validity, and construct validity (Sudaryono, 2012).

Quantitative Analysis (Empirical Validity)

Quantitative analysis known as empirical validity is conducted after the test items are tested to representative samples to determine the effectiveness and usefulness of the test items. In this type of item analysis, quantitative analysis or empirical validity emphasizes on analyzing the internal characteristics of the test through statistical data after the representative samples' responses of each test item are scored (Sudaryono, 2012).

Surapranata (2009) stated that one of the purposes of quantitative analysis is to increase the quality of the test. After conducting a quantitative analysis, the extent of the quality of an item can be determined whether the item is *accepted*, *revised*, or *eliminated*.

a. *Acceptable*. The test items are acceptable when it is proven through empirical judgment that they are effective to distinguish among students and it is already supported by adequate statistical data.
b. *Revised*. The test items are revised when it is proven through empirical judgment that there are some weaknesses on the test items.
c. *Eliminated*. The test items are eliminated when it is proven through empirical judgment that they are not useful.

The internal characteristics determined through quantitative analysis are intended to cover item difficulty, item discrimination, and item distracter.

1. Item of Difficulty

Item difficulty is the proportions of the students who responded correctly to an item. Item difficulty which is commonly known as *p*-value refers to the percentage of test-takers who responded to an item correctly (Sabri, 2013). To know an index of item difficulty (*P*), it can be determined by calculating the proportion of test takers who answer the item correctly. The formula to calculate the item difficulty index is as followed (Kunandar, 2013).

The Formula of Item Difficulty Index

$$P = \frac{\mathrm{B}}{\mathrm{T}}$$

In which:

P: Index of item difficulty
B: Numbers of test takers in the total group who pass the item
T: Total numbers of test takers in the group.

After finding out the index of item difficulty, the index is used to determine the item difficulty level. The item difficulty level determines whether an item is considered *difficult, moderate,* or *easy* according to the range scale of the index. The classification of the item difficulty level is as followed (Kunandar, 2013).

The Classification of Difficulty Item Level

P	**Difficulty Level**
0.00 – 0.30	Difficult
0.31 – 0.70	Moderate
0.71 – 1.00	Easy

2. Item of Discrimination

Item discrimination is a measure intended to distinguish between the performance of the students in the high score group and students in the low score group, and item discrimination is determined based on the discrimination index. Sabri (2013) suggests that the discrimination index fundamentally distinguishes students who are knowledgeable and students who are not knowledgeable, revealing the score results of top scorers and low scorers in each item. In this case, the discrimination index determines the ability of an item to distinguish among the students from how well students know the materials have been tested.

Item discrimination index (D) can be obtained by dividing the test takers into three groups according to their scores on the test as a whole: an upper group consisting of the 27% who make the highest score, a middle group consisting of 46%, and a lower group consisting of the 27% who make the lowest score. The following formula is employed to determine the item discrimination index (Kunandar, 2013).

The Formula of Item Discrimination Index

$$\mathrm{D} = \frac{2\ (\mathrm{A} - \mathrm{B})}{\mathrm{T}}$$

In which
D: Item discrimination index
A: Numbers of test takers in the upper group who pass the item
B: Numbers of test takers in the lower group who pass the item
T: Total numbers of test takers in the group

After finding the index of an item discrimination index, the discriminating power can be determined. To determine a discriminating power, classification is used to indicate whether the extent of the quality of each test item is considered as *excellent*, *good*, *satisfactory*, *poor*, or *very poor* (Sudijono, 2011).

The Classification of Discriminating Power

Discrimination Index	**Quality**
0.71 – 1.00	Excellent
0.41 – 0.70	Good
0.21 – 0.40	Satisfactory
≤ 0.20	Poor
Negative Value on D	Very Poor

Ideally, students who know the content and who perform well on the test overall should be the ones who know the contents. Otherwise, problems will arise if students getting the correct answer on the test do not know the contents being tested. Concerning this case, the negative value on discrimination index is addressed. Surapranata (2009) stated that theoretically, the negative value on discrimination index indicates that test-takers who are less knowledgeable are able to respond correctly to an item than those who are not. In shorts, items with a negative value on its discrimination index show the quality of the test takers upside down.

3. Item of Distractor

In multiple-choice testing, the intended option is called the ‘key,' and each incorrect option is called a ‘distractor’ (Fulcher, 2007). In a good test, the distractor is more likely to be chosen equally by students who responded to the test item incorrectly. On the other hand, in a poor test, the distractor is chosen unequally. A distractor is considered a good distractor when the total numbers of test takers choose the same distractor. Distractor index is calculated by employing the following formula (Arifin, 2013).

The Formula of Distractor Index

$$\mathrm{IP} = \frac{P}{(\mathrm{N}-\mathrm{B})/(\mathrm{n}-1)}x100$$

In which:

IP: Distractor index
P: Number of students choosing distractor
N: Number of students taking the test

The Relationship Between Item Difficulty Level and Item Discriminating Power

An item in a test should neither be too easy nor too difficult. Concerning accuracy in distinguishing between students who are knowledgeable (top scorer) and those who are less knowledgeable (lower scorer), the level of item difficulty directly influences to item discriminating power. When everybody chooses the correct answer (*P* = 1), or everybody gets the item (*P* = 0), it means that the item cannot be used to differentiate the upper group students and the lower group students because the *P*-value is too extreme. The relationship between item difficulty index and the index of item discriminating power is illustrated in the table as followed (Surapranata, 2009).

The Maximum Index of Item Difficulty Functioning the Index of Item Discriminating Power

***P* Value**	**D Maximum**
1.00	0. 00
0.90	0.20
0.80	0.40
0.70	0.60
0.60	0.80
0.50	1. 00
0.40	0.80
0.30	0.60
0.20	0.40
0.10	0.20
0. 00	0. 00

Based on the above table, it shows that item difficulty (P) = 0.50 obtains a maximum index of item discrimination (D) = 1.00. It indicates that an item which has 0.50 in its item difficulty index has the best item discriminating power. As a result, the item difficulty level is used as an indicator to determine the accuracy to differentiate among students. If an item is very easy or very hard, the item is not likely to be very discriminating. In other words, a very easy or very difficult item is not a good discriminator. If an item is so easy that nearly everyone gets it correct, or so difficult to the extent that nearly everyone gets it wrong, then it becomes hard to discriminate those who actually know the content of the test from those who do not. Based on this relationship between item difficulty level and item discriminating power, the extent of the quality of the English mid-term test items is determined.

METHOD

The objectives of this study were to know and to describe the extent of the quality of the English test items concerning difficulty level and discriminating power of multiple choice items made by an English teacher. Later, the extent of the quality of each test item can be determined. This study was designed as a descriptive study and conducted with 171 respondents of students at MAN 1 Kota Tangerang Selatan 2017/2018 academic year. In this study, one hundred and seventy-one students were split into lower, middle, and upper group based on their score after doing the test. In the analysis process, 30 students out of 171 were taken from 27% of the upper (15 students) and 27% of the lower group students (15 students). According to Crocker, taking 27% from the upper and 27% from the lower group is a widely used technique to divide the group for determining item discrimination because 27% is the most stable and sensitive percentage (Surapranata, 2009).

A test of 30 items was used for data collection. In this study, data were gathered from the students’ response to each test item in the form of multiple choices. The data were gathered from thirty of students' answer to items in the English test. The test was administered for the mid-term test. To know an index of item difficulty (P), it can be determined by calculating the proportion of test takers who answer the item correctly. The formula to calculate the item difficulty index is as followed (Kunandar, 2013).

The Formula of Item Difficulty Index

$$P = \frac{\mathrm{B}}{\mathrm{T}}$$

In which:

P: Index of item difficulty
B: Numbers of test takers in the total group who pass the item
T: Total numbers of test takers in the group.

After finding out the index of item difficulty, the index is used to determine the item difficulty level. The item difficulty level determines whether an item is considered *difficult*, *moderate*, or *easy* according to the range scale of the index. The classification of the item difficulty level is as followed (Kunandar, 2013).

The Classification of Item Difficulty Level

P	**Difficulty Level**
0.00 – 0.30	Difficult
0.31 – 0.70	Moderate
0.71 – 1.00	Easy

Finally, item discrimination is a measure intended to distinguish between the performance of the students in the high score group and students in the low score group, and item discrimination is determined based on the discrimination index. Sabri (2013) suggests that the discrimination index fundamentally distinguishes students who are knowledgeable and students who are not knowledgeable, revealing the score results of top scorers and low scorers in each item. In this case, the discrimination index determines the ability of an item to distinguish among the students from how well students know the materials have been tested.

Item discrimination index (D) can be obtained by dividing the test takers into three groups according to their scores on the test as a whole: an upper group consisting of the 27% who make the highest score, a middle group consisting of 46%, and a lower group consisting of the 27% who make the lowest score. The following formula is employed to determine the item discrimination index (Kunandar, 2013).

The Formula of Discrimination Item Index

$$\mathrm{D} = \frac{2\,(\mathrm{A} - \mathrm{B})}{\mathrm{T}}$$

In which
D: Item discrimination index
A: Numbers of test takers in the upper group who pass the item
B: Numbers of test takers in the lower group who pass the item
T: Total numbers of test takers in the group

After finding the index of an item discrimination index, the discriminating power can be determined. To determine the discriminating power, classification is used to indicate whether the extent of the quality of each test item is considered as *excellent*, *good*, *satisfactory*, *poor*, or *very poor* (Sudijono, 2011).

The Classification of Discriminating Power

Discrimination Index	**Quality**
0.71 – 1.00	Excellent
0.41 – 0.70	Good
0.21 – 0.40	Satisfactory
≤ 0.20	Poor
Negative Value on D	Very Poor

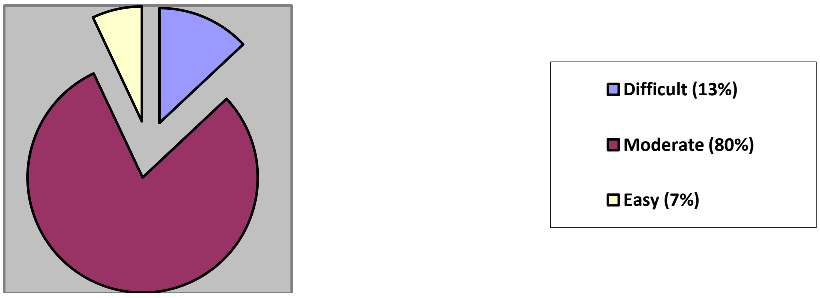
Total scores of the students were entered in Microsoft Excel sheet and arranged in descending order, then 30 (27%) upper and lower group students were selected for item analysis. The middle group students (46%) were excluded from the analysis as suggested in the literature. In this study, *Anates program version 4.0.2* was also employed to describe the data. The formulae for difficulty level and discriminating power discussed above were used for analysis.

FINDINGS AND DISCUSSIONS

Findings

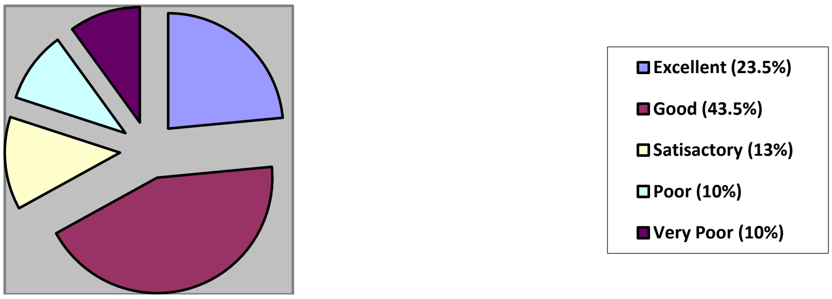
Regarding item difficulty level, there are 4 (13%) items fall into *difficult*, 24 (80%) items fall into *moderate*, and only 2 (7%) items fall into *easy*. Above all, to make the information easier to read, the following is the chart of the percentage of item difficulty level.

Chart 1. The percentage of difficulty level



Meanwhile, in the discrimination index it is found that 7 (23.5%) items have *excellent* quality, 13 (43.5%) items have *good* quality, 4 (13%) items have *satisfactory* quality, 3 (10%) items have *poor* quality, and 3 (10%) items have *very poor* quality or in the negative value on discrimination index. To make the information easier to read, the following is the chart of the percentage of item discriminating power.

The Percentage of Item Discriminating Power



Discussion

Concerning on the results of the study as presented in chart 1 and chart 2 and with the theories as presented in the literature review, it is essential to bear in mind that there is a relationship between the index of item difficulty and its discrimination index. Theoretically, when everyone chooses the correct answer (P = 1), or everyone chooses the item (P = 0), then the item cannot be used to distinguish the upper and the lower group because the P-value is too extreme. When an item is either very easy or very hard, it is not likely to be very discriminating. In other words, a very easy or very difficult item is not a good discriminator. When an item is so easy that nearly everyone gets it correct or so difficult to the extent that nearly everyone gets it wrong, then, it becomes very hard to discriminate those who actually know the content of the test from those who do not.

Also, it is typically recommended that the item discrimination index be at least 0.20, and it is best to aim even higher. Thus, the negative value on discrimination index must be addressed. Theoretically, items with a negative discrimination indicate that either the students who performed poorly on the test overall got the item correct, or that students with high overall test performance did not get the item correct. In other words, the negative value on discrimination index could signal some problems, such as a mistake on the scoring key, poorly prepared students were guessing correctly, or well-prepared students were somehow justifying the wrong answers.

Finally, the extent of the quality of an item can be determined whether it is *acceptable, revised,* or *eliminated* as proposed by Surapranata (2009). First, the test items are *acceptable* when it is proven through empirical judgment that the items are effective to distinguish among students and it is already supported by adequate statistical data. Then, the test items are *revised* when it is proven through empirical judgment that there are some weaknesses on the test items. Lastly, the test items are *eliminated* when it is proven through empirical judgment that they are not useful to differentiate students who are knowledgeable from those who are not knowledgeable.

Based on the result from the study, it shows that twenty-four items out of 30 (80%) are accepted in *excellent, good, and satisfactory* quality, and it is already proven by adequate statistical data that the items are effective to differentiate among students. Three items (10%) have *poor* quality, and they are needed to be revised to improve the quality of the test items. Finally, there are three items (10%) have very poor quality. *Very poor* quality means that the items are in the negative value on discrimination index. Therefore, they are discarded since it is proven by statistical data that they fail to distinguish the upper group and the lower group students.

CONCLUSION AND SUGGESTION

The major implication of this work is the realization that item analysis is essential for making a good test, and improving test items. By conducting item analysis, the extent of quality of a test can be determined. Also, the results of item analysis are intended to find out to what extent the teacher-made tests clearly show the difference among students concerning their level of knowledge of content and material of lessons being tested. In this study, the findings showed that the English test has 24 acceptable items (80%) from the quality *excellent, good, and satisfactory* and it is proven by these adequate statistical data that the items are effective to distinguish among students. Three items (10%) have *poor* quality, and they can be revised to improve the quality of the test items. Lastly, there are three items (10%) have *very poor* quality, or in the negative value on discrimination index to the extent they are eliminated. The significances in this study go to the English teacher as the test maker and test developer of the English mid-term test items.

Based on the conclusion of the research, some suggestions are delivered to the test maker that the statistical results should be used along with the item content to determine what should be done to improve the item tests. Items in the *excellent, good, and satisfactory* quality can be kept in items banks to be reused in the future. Items in *poor* quality may still be usable after modest changes are made to improve the items. The items with negative discrimination index could signal some problems, for example, a mistake on the scoring key, poorly prepared students guessed correctly, or well-prepared students were somehow justifying the wrong answers. Therefore, the items are needed to be eliminated. Also, it is needed to be ensured there is only one possible answer, the question is written clearly, and the answer key is correct. Finally, this work can be repeated in any other subjects to create a good test and improve test items to which the extent of quality of a test can be determined.

REFERENCES

Arifin, Z. (2013). *Evaluasi Pembelajaran: Prinsip, Teknik, Prosedur.* Bandung: Remaja Rosdikarya.

Boopathiraj, C. a. (2013). Analysis of Test Items on Difficulty Level and Discriminating Index in the Test for Research in Education. *International Journal of Social Sciences & Interdisciplinary Research, 2* (2), 189-193.

Desheng, C. a. (2013). Testing and Evaluation of Language Skills. *Journal of Research & Method in Education, 1* (2), 31-33.

Education, C. D. (2004). *Key Elements of Testing.* Department of Education State of California.

Foyewa. (2015). Testing and Evaluation in English Language Teaching - A Case of O Level English in Nigeria. *International Journal of English Language Teaching, 3* (6), 32-40.

Fulcher, G. (2007). *Language Testing and Assessment: An Advanced Resource Book.* New York: Routledge.

Kunandar. (2013). *Penilaian Autentik (Penilaian Hasil Belajar Peserta Didik Berdasarkan Kurikulum 2013): Suatu Pendekatan Praktis.* Jakarta: Raja Grafindo.

Quaigrain, K. a. (2017). Using Reliability and Item Analysis to Evaluate A Teacher developed Test in Educational Measurement and Evaluation. *Research Article, 4* (1), 1-11.

Sabri, S. (2013). Item Analysis of Student Comprehensive Test for Research in Teaching Beginner String Ensemble Using Model-Based Teaching Among Music Students in Public Universities. *International Journal of Education and Research, 1* (12), 241-254.

Sudaryono. (2012). *Dasar-Dasar Evaluasi Pembelajaran.* Yogyakarta: Graha Ilmu.

Sudijono, A. (2011). *Pengantar Evaluasi Pendidikan.* Jakarta: Grafindo Persada.

Sujana, I. M. (2000). Movements in Language Testing: From Grammar-Based to Communicative Language Testing. *Jurnal Ilmu Pendidikan FKIP UNRAM, 13* (48), 1-8.

Surapranata, S. (2009). *Analisis, Validitas, Reliabilitas, dan Interpretasi Hasil Tes.* Bandung: Remaja Rosdakarya.

The Ministry of National Education Republic of Indonesia. (2003). *Act of the Republic of Indonesia Number 20 on National Education.*

Young, J. W. (2013). *Guidelines for Best Test Development Practices to Ensure Validity and Fairness for International English Language Proficiency Assessment.* Educational Testing Service.