QUALITY ANALYSIS OF ENGLISH FINAL TEST FOR LIGHT VEHICLE ENGINEERING STUDENTS AT SMK NEGERI 1 NGULING

Maya Fitriyah Firdaus
Universitas PGRI Wiranegara, Pasuruan
mayafirdaus75@gmail.com

Lestari Setyowati
Universitas Negeri Malang, Malang
lestari.setyowati.fs@um.ac.id

Barotun Mabaroh
Universitas PGRI Wiranegara, Pasuruan
barotunmabaroh@yahoo.com

Abstract: The purpose of this study is to investigate the quality of an English final test of SMKN 1 Nguling seen from its content validity, reliability, item difficulty, discriminating power, and effectiveness of distractor. The design of this study was qualitative research which focuses on content analysis and descriptive quantitative. The result shows that the test was considered acceptable but it needs revisions to increase the reliability of the test. In particular, the researchers found out that the content validity of the test is very high, and the reliability of the test is moderate (r. 0.427). However, the test had poor item difficulty as half of the test items test was in the difficult category. The test also has poor discriminating power. Finally, in terms of the effectiveness of distractors, the result shows that 24 items need some revisions. The researchers concluded that even though the test is high in content validity and moderate in its reliability, the test still needs some improvements, especially in discriminating power and item distracter.

Keywords: *English Final Test, Item analysis, Test Quality*

One of the pedagogic competencies that a teacher must have is to evaluate the result of the teaching and learning process. An assessment can give a lot of information about the teaching and learning process, both its success and its failure (Sulistyo, 2017). Sulistyo (2017) also states that an assessment does not serve as a judgment, rather as a guide about the teaching and learning process. According to Kellaghan, T., & Greaney, V. (2001) the purpose of the assessment is to provide information about the failure and success of the teaching and learning process in a particular institution and to know how good the students do in classroom learning. Itcan provide information for the publisher about the effectiveness of a particular book used in the teaching and learning process.

The teacher as a test maker had to know how to make a good test. A test must be practical, valid, and reliable (Brown, 2001: 385; Arikunto, 2008). Further, Arikunto (2008) states that a characteristic of a good test is it has a good level of difficulty and discriminating power. The validity and reliability of the test can be seen from the result of measuring the test. Brown and Lee (2015: 492) said that the most complex character of a good test is its validity. Farida (2017: 159) also said that content validity is the validity of assessment instrument which represent the whole materials. Thus, validity, in brief, is to what extent the test measures what it is supposed to measure. The test must aim to provide a true measure of the skill which it is intended to measure. Validity is the most important characteristic of a test since it concerns test quality. Validity measures what is supposed to measure and gives the teacher/the headmaster/the test maker the information about what they want to know (Sulistyo, 2018). If the test is invalid, they will be misled by the information provided by the result of the test. The higher the degree of the validity, the better the test will be.

Yet, a good test should not only be valid but also must be reliable (Cohen & Swerdlik, 2010). According to Sulistyo (2018: 43), a reliable test is demonstrated by the scoring consistency within raters or inter-raters, between raters or inter-raters, and across time and place. Further, Arikunto (2003: 87) stated that a reliable measure provides a consistent and stable indication of the characteristic being investigated. It can be concluded that reliability is the consistency or the stability of test scores. A typical way to find out the reliability (its strength and direction) of a test is by correlating two or more sets of scores by using statistical software (Luoma, 2004). The relationship between the set of scores is

known as the correlation coefficient (Davies et al., 1999). The lowest correlation coefficient is 0.00 while the highest is 1.00. The values which are close to 0 indicate no relationship, on the other hand, the value that is close to 1 has a perfect correlation (Hughes, 2003). Yet, neither extreme values come into practice.

Thus, a test needs to have a sound quality because a good test would provide the right information for the teacher in making an accurate decision about the students' performance. Unfortunately, some research found that many teacher-made tests do not show satisfying quality (Karim, Sudiro, & Sakinah, 2021; Manalu, 2019; Hartati, & Yogi, 2019, Setyawati, Putri, Pusparini, 2018) In spite of this fact, there are research on the quality of the final test at the High School level in Pasuruan showing that the tests made by the teacher or the team teacher in Pasuruan are good. Firstly, Aulia (2017) investigated an item analysis of the English final test for the twelfth grade in SMA Negeri 1 Kejayan. The result shows that the test has high content validity because there are 33 items(73,3%) that are appropriate to the base competence of the syllabus. But the reliability of the multiple-choice test was low. In terms of the item difficulty, the test is not difficult nor too easy for the students. The finding also reveals that the item distractors of the test were not satisfactory. This means that the test has many deficient item distractors. The second study was conducted by Wahyuningrum (2017). She investigated an item analysis of English try out for the national final examination at MA KHA. Wahid HasyimBangil. The finding showed that the content validity of the test is very high because 100% of the item test was appropriate with the syllabus. The reliability of the test was moderate, the coefficient of the degree reliability is 0,5304. Then, the item difficulty was also moderate.

So far, there is scarce research on the quality of a test in vocational high schools, especially in the Pasuruan region. Most of the test item analysis are done in high schools and the information about the test quality in vocational high school is limited. Thus, more studies need to be conducted to find out the test quality either in the state vocational high schools or in the private ones. Investigating the test quality will give valuable information for the teachers, the headmaster, and the curriculum developer about the strengths and weaknesses of the teaching and learning process. One of the state's vocational high schools in Pasuruan is SMKN I Nguling. SMKN 1 Nguling is one of the favourites vocational high schools in Pasuruan region. It has a lot of achievements, not only in the academic setting, but also the non-academic ones. Among others, the students of SMKN 1 Nguling has won the runner-up position in the story telling contest in the north regions of East Java in 2019. The school is also actively participated in many English competitions held by the universities in East Java. In this research, the researchers focused on the light vehicle engineering department (*Teknik Kendaraan Ringan*/TKR) classes Based on the preliminary study, many students felt that the test made by the teachers are difficult. Many of them got low scores in English test. The researchers suspect that the English test made by the teachers are too difficult for the students. Based on the ground of the discussion and the preliminary study, the researchers are intended to investigate the quality of the final test in this school in terms of its validity, reliability, item difficulty, discriminating power, the effectiveness of distractor of the English final test for second grade of SMK Negeri 1 Nguling.

METHOD

This study used descriptive quantitative study. Descriptive quantitative research is a type of research that describe and interpret the existing phenomena without any attempt to manipulate individuals or situations. The quantitative approach is used because mostly the data are in the numerical form. The subject of this study was all second-grade students of SMK Negeri 1 Nguling who major in light vehicle engineering (*Teknik Kendaraan Ringan*/TKR). In all, there were 101 students. Furthermore, the object of this study was the English final test 2018/2019 for the second grade of SMK Negeri 1 Nguling. The final test of SMK Negeri 1 Nguling was made by a team of teachers of the school. The English team teachers in SMKN 1 Nguling develops the final test. There are 35 items in the test. Thirty items of the test use multiple choice items, while the last five items are subjective test type. As item analysis works well with the multiple choice items, the researchers decided to take the multiple-choice ones only, and leaving the subjective type untouched for analysis.

There were 2 kinds of instruments used. They were documentation and human instrument. The researchers documented the English final test, answer key, students answer sheet and syllabus. Then, a human instrument was used to collect and analyse the data. The test was done on May, 15th 2019. The researchers used data codification to categorize and analyse the data. The codification applies to the

student's identity and the test category. In analyzing content validity, the researchers collected it through the following step: (1) Made a list of standard competencies, basic competencies, and indicators for second grade of SMK Negeri 1 Nguling. (2) Analysed the test item and standard competencies of the syllabus whether those terms are covered by the final test. (3) Counted the percentage of the test item. (4) Concluded the result of the analysis. Meanwhile, to analyse the reliability, the researchers used Cronbach's Alpha. There were several steps to determine the reliability of the test: (1) Made tabulation of test score for each test item. (2) Measured reliability of a test by using Cronbach alpha in SPSS 23 version. (3) Classified the result into criteria of reliability, namely very high (0,81 – 1,00), high (0,61 – 0,80), moderate (0,41 – 0,60), low (0,21 – 0,40), and very low (0,00 – 0,20) (Sudijono, 2008). To find out the item difficulty, the researchers used the index difficulty.

In analysing discriminating power, the researchers used the formal proposed by Heaton (1988). The steps to analyse discriminating power are: (1) Arrange the script in rank order to rank total scores and divided it into three groups: upper group, middle group, and lower group. (2) Classified the upper group of the sample is 27% students who got highest grades from the whole sample, and the lower group is 27% students who got the lowest group from the whole sample. The rest students are categorized as the middle group. (3) Counting the number of those candidates in the upper group answering the first item correctly, and the account number of the lower group answering the item correctly. (4) Subtract the number of the correct answer of the lower group from the number of the correct answer in the upper group: i.e find the difference in the proportion passing in the upper group and the proportion passing in the lower group.

FINDINGS AND DISCUSSION

The purpose of this study is to investigate the quality of an English final test of SMKN 1 Nguling. The test was the teacher made-test. This is crucial since this research result would support the better measurement of whether the learning goal is successfully reached or not through the process. This is due to Brown's statement (2001:384) that a test is a method of measuring a person's ability or knowledge in each domain.In addition, Sudijono (2011:67) and Sudjana (2014) also state that a test as an assessment device was a set of items or questions given to the students in the form of spoken, written, or performed.

By testing teacher and the headmaster will get information about the failure and success of the teaching and learning process in a particular institution. It can also give information for parents about how good their children do in classroom learning.

Content Validity of English Final Test

The test could be said a good test when it has the certain characteristic of a good test. Brown (2001) defines there were three criteria of a good test, they are validity, reliability, and practicality. Similarly, Sulistyo (2015) also explains that a good test must meet the requirements of elicitation tools or instruments which are reliability, validity, practically/usability, and economy.

Validity had different types. But they had a similar classification of validity. According to Brown (2015), there are three types of validity, namely content validity, face validity, and construct validity. Farhady (2012:38) states that content validity refers to the correspondence between the content of the test and the content of the materials to be tested. He also adds the content of the test should be a reasonable and representative sample of the total content to be tested. Thus, the items should be representative enough of each portion of the analysis or outline. Meanwhile, face validity refers to 'the face' of the test. Sulistyo (2015:62) states that if the test looks are intended to measure what is intended to measure, the test can be said to have face validity. Thus, it simply means the way the test *looks*-- to the examinees, test administrators, educators, and the like. If the test appears irrelevant, or inappropriate, knowledgeable administrators will hesitate to adopt the test and examinees will lack the proper motivation. And the last is the construct Validity. Construct is a concept that can be observable and measurable (Arifin, 2009:257). Further, Akbari (2012:32) said that construct validity is concerned with the psychological reality of a test, He states that construct validity asks the question of what it means to know the language and what the nature of that knowledge. In short, construct validity refers to the capability of measuring certain specific characteristics following a theory of language behaviour and learning.

The researchers used the current curriculum to analyse the content validity. The result shows that all test items were appropriate with the syllabus. But, not all basic competencies were covered in

the test items. There were just two basic competencies covered in the test item. They are about the factual report and analytical exposition text. Moreover, the test gives more focus only on reading and vocabulary instead of other skills.

Table 1. The Reading Text for Test Items with Basic Competence 4.13

A thunderstorm is a form of weather characterized by lightning and thunder. Over 40.000 thunderstorms occur throughout the world each day. Basically, however, there are two main types of thunderstorm: ordinary and severe. Ordinary thunderstorms are common summer storms which last about one hour. They are usually accompanied by rain and occasionally small hail. An ordinary thunderstorm cloud can grow up to 12 kilograms high. Severe thunderstorms are really dangerous. They are capable of producing baseball-sized hail, strong winds, intense rain, flash floods, and tornadoes. They can last several hours and can grow 18 kilometers high.
1. What is the thunderstorm? a. A form of weather characterized by lightning and thunder b. Common storms which last about one hour c. Really dangerous d. Producing baseball-sized hail

Item number 1 is an example of a reading skill about factual report which is appropriate with basic competence 4.13, that is capturing the meaning (determining or looking for explicit or implicit meanings in the text) in factual scientific texts (factual reports), oral and written, simple, about people, animals, objects, natural and social symptoms, and events, related to other subjects in Class XI. The other test items were also made to measure the students' achievement on different basic competence. The summary of the appropriateness is presented in table 1.

Table 2. The Number of Related Items with the Syllabus

Standard competence	Related item	Total	Percentage
3.9 4.13	Reading: 21 Reading: 1, 2, 3, 4, 5, 6, 7, 8, 14, 15, 16, 17, 18, 19, 20	16	54%
3.10 4.14	Reading: 11, 12, 23 Reading: 9, 10, 13, 22, 24 Vocabulary: 25, 26, 27, 28, 29, 30	14	46%
Total		30	100%

Based on the data, it can be said that the English Final Test for the second grade of SMK Negeri 1 Nguling has 100% of appropriateness with the syllabus. This means that the test has very high content validity. The test has 24 items for reading skills which discuss factual report text and analytical exposition. While there were 6 items for vocabulary skill which are also about the analytical exposition. Therefore, it could be said that the test was dominated by the material about factual report text and analytical exposition.

Reliability of English Final Test

The next characteristic of a good test is reliability. Ngalimun (2018: 262) states that a reliable measure in one that provided a consistent and stable indication of the characteristic being investigated. Further, Kemp, Morrison, & Ross (1994:1 67) argues that reliability refers to a test's ability to produce consistent results whenever used. He also adds if the same learners without changes in the preparation were to take the same test or an equal form of the test, there should be little variation in the score. In other words, a reliable test should be demonstrated by the scoring consistency within raters or inter-rates, between raters or inter-raters, and across time and place (Sulistyo, 2015: 43). In this present study, to find out the reliability of the test, the researchers used Cronbach's Alpha in SPSS 23.

The table shows that the test reliability was 0.427. This means that the test had moderate reliability because of the test on the criteria between 0,41-0,60. Sudijono (2008) states that any reliability coefficient between 0.41-0.60 belongs to moderate reliability. Reliability is one of the characteristics of a good test (Brown, 2001:386). To improve the reliability of the test, the test makers can revise some items. To improve the reliability of the test, as stated by Wells & Wollack (2003) is by adding some test items or improving test length.

Item Difficulty of English Final Test

Item difficulty shows how difficult or easy the test items are. Theoretically, test items should not be too easy or too difficult for the test takers. From the analysis, the result reveals that there were 16 difficult items (number 9, 10, 11, 12, 14, 17, 19, 20, 21, 22, 23, 24, 25, 27, 29, 30); eight moderate items (number: 6, 7, 8, 13, 15, 16, 26, 28) and six easy items (number: 1, 2, 3, 4, 5, 18).

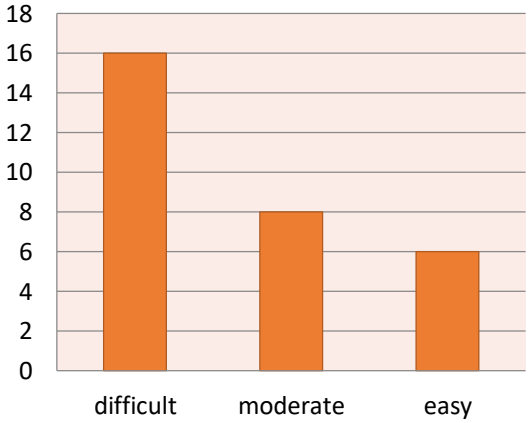

Figure 1. Item difficulty

Sudjana (2014:135) and Arifin (2016) states that a good test had to be balanced between difficult, moderate, and easy test items. He gives an example that if there were 60 items, 20 for difficult items, 20 for moderate items, and 20 for easy test items. The data of the study shows that the English final test was not good in terms of its item difficulty. Figure 1 shows the item difficulties are not in a balanced condition. Since most of the test items fall in the difficult category, the researchers believe that the test makers should revise the option in some test items.

Discriminating Power of English Final Test

Item analysis aims to identify whether the test is good or not. Test item analysis is a process or procedure which can be used by a teacher to find out the quality of a test (Basuki & Hariyanto, 2017:129). Arikunto (2016:222) argues that the items that must be identified in the test are item difficulty, discriminating power, and effectiveness of distractor. A good test should not be too easy, nor too difficult. Arikunto (2016) states that there were three classifications on the level of item difficulty namely easy (0,71 – 1,00), moderate (0,31 – 0,70), and difficult (0,00 – 0,30). The next item in test analysis is the discriminating power. Discriminating power was the capability of a test to distinguish the students who were in the upper group and who was in the lower group.

Arikunto (2016: 226) states that if the upper and the lower group can answer with the correct response, the test is not good because there is no discriminating power. Likewise, if both of group gives a false response, it is also not a good test. To know the level of discriminating power in the test, Arikunto (2016:232) offers four classifications of discriminating power. They are poor, satisfactory, good, and excellent. The next element is the effectiveness of the distractor. Distractor or item distractor was commonly used in a multiple-choice test. Sulistyo (2015: 224) said that the quality of a multiple-choice item depends on the quality of the distractor. In this present study, to analyse the discriminating power, the researchers classified the scores into two; the upper and lower group. To measure the upper and lower group, the researchers took 27% of the total student for the upper group and 27% for the lower group (Arifin, 2016). In this study, the total number of students was 101, so there were 27 students for the upper group and 27 students for the lower group. However, the researchers added similar scores for the upper and lower group. So, there were 29 scores in the upper group and 31 scores in the lower group.

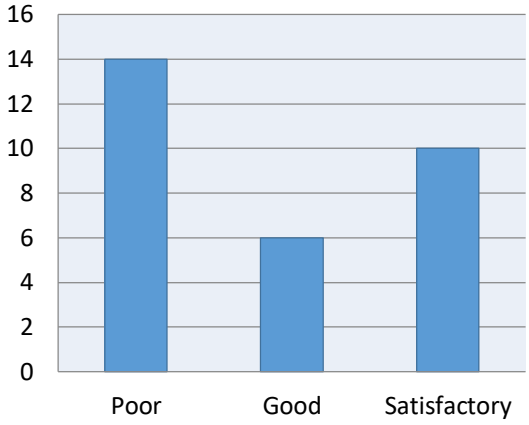

Figure 2. Discriminating power

The result shows that there were 14 poor items (number 1, 2, 3, 8, 13, 14, 16, 18, 22, 23, 24, 26, 28, 29, 30). The good items are only six (number 5, 7, 17, 19, 20, 21). And ten satisfactory items (number 4, 6, 9, 10, 11, 12, 15, 24, 25, 27). Discriminating power was the capability of a test to distinguish the students who were in the upper group and who was in the lower group (Arikunto, 2016). He further states that if the upper and the lower group can answer the test items with the correct responses, it can be said that the test is not good because it has no discriminating power. Likewise, when all groups give false responses, the test is also considered not a good test. The data shows that There were no excellent criteria in the test. The test items could be said in the excellent category if most of the upper group choose the right option, and most of the lower group choose the right option. Looking at the data, the test maker had to revise some test items if the same test should be reused next time.

Effectiveness of Distractor

To determine the effectiveness of item distractors in the English final test, the researchers used Arifin's criteria (2016). In interpreting the effectiveness of distractor, the researchers conclude that the right answer of six test items (item number 1, 2, 3, 4, 5, 18) are not selected by any students. While the incorrect answer of the eight test items (number 17, 19, 24, 26, 27, 28, 29, 30) is chosen by a lot of participants. In addition, the incorrect options of test items number (7, 10, 11, 12, 13, 14, 20, 21, 22, 23) are chosen by most of the test takers. Only six items (number 4, 6, 8, 13, 15, 16) have a balanced distribution. Thus, it can be concluded that 24 items should be revised, and 6 items should be maintained. Multiple-choice items should have good item distractors (Sulistyo, 2015:224). Based on the quality of the distractors, the test under study is not good. Thus, the test maker must revise some items in the test and improve the quality of the distractor. Gierl, et al (2017) state that there are two strategies in distractor development. The first strategy focuses on a list of plausible but incorrect options linked to common misconception, and the second strategy focuses on the similarity by creating distractors that are similar in content and structure relative to the correct option.

The result of this study implies that the teacher-made test, either made individually or made by a team, needs to be developed carefully. Previous research have shown that many teacher-made tests do not show satisfying quality (Karim, Sudiro, & Sakinah, 2021; Manalu, 2019; Hartati, & Yogi, 2019, Setyawati, Putri, Pusparini, 2018). Thus, this study support previous studies in terms of the quality of the teacher-made test. Even though the test has high content validity and moderate reliability, it does not mean that the test is in good quality. The result of the present study reveals that the English final test of the second graders majoring in light vehicle engineering needs some improvements in terms of item difficulties, discriminating power, and item distractors.

CONCLUSION

The Analysis in this research concludes that the English final test for the second grade of SMK Negeri 1 Nguling 2018/2019 academic year is considered acceptable and needs some improvements to upgrade its quality. First, based on the analysis of content validity, the final test of the second-grade students of SMK Negeri 1 Nguling had very high content validity as all items represent the basic competence of the syllabus. Second, the reliability of the test was in the moderate category (r. 0,427). In terms of the item difficulty of the test was not good. Half of the test items were too difficult. Fourth, in

terms of the discriminating power, most of the test items are in the poor category. And finally, in terms of the effectiveness of distractors, the study found that almost two-thirds of the options in the multiple-choice test items are not good distractors. In short, the test needs some revisions.

The researchers address the suggestions to the test makers and the future researchers. First, the test makers of the English final test of the second-grade students in SMK Negeri 1 Nguling revise the test, in terms of the test items and its distractors. The test makers should also improve their knowledge and practical skills on how to make a good test, such as by joining seminars and workshops in language testing for classroom application. Secondly, not much research is dedicated to the language used in the test. The difficulties faced by the students in doing the test might be caused by the wording of the questions or instructions. Thus, future researchers can investigate the quality of the language used in the test for any level of students, more particularly, the vocational school students.

REFERENCES

Akbari, Ramin. 2012. Validity in Language Testing. Coombe,Christine (eds), *The Cambridge Guide to Second Language Assessment* (pp.32). USA: Cambridge University Press.

Arifin, Z. 2009. *Evaluasi Pembelajaran.* Bandung: PT. Remaja Rosdakarya

Arifin, Z. 2016. *Evaluasi Pembelajaran.* Bandung: PT. Remaja Rosdakarya Offset

Arikunto, S, 2003, Dasar-Dasar Evaluasi Pendidikan, edisi revisi, Bumi Aksara, Yogyakarta

Arikunto, S. 2008. Prosedur Penelitian Suatu Pendekatan Praktik. Jakarta: Rineka Karya.

Arikunto, S. 2016. *Dasar-Dasar Evaluasi Pendidikan.* Jakarta: Bumi Aksara.

Aulia, Nurimah Arum. 2017. *An Item Analysis of English Final Test for The Twelfth Grade in SMA Negeri 1 Kejayan.* Unpublished. S1 Thesis. Pasuruan: English Educational Study Program STKIP PGRI Pasuruan.

Basuki, Ismet & Hariyanto. 2017. *Asesmen Pembelajaran.* Bandung: PT. Remaja Roosdakarya

Brown, H. D. & Lee, H. K 2015. *Teaching by Principle: An interactive Approach to Language Pedagogy Fourth Edition.* San Francisco: Pearson.

Brown, H. D. 2001. *Teaching by Principle: An interactive Approach to Language Pedagogy Second Edition.* San Francisco: Longman

Cohen, J.R. & Swerdlik, M.E. 2010. *Psychological Testing and Assessment an Introductional to Test and Measurement Seventh Edition.* New York: McGraw-Hill

Davies, A., Brown, A., Elder, C.& Hill, K. (1999). *Dictionary of Language Testing.* Cambridge: Cambridge University Press.

Farhady, H. 2012. Principle of Language Assessment. Coombe, Christine (eds), *The Cambridge Guide to Second Language Assessment* (pp.38). USA: Cambridge University Press.

Farida, I. 2017. *Evaluasi Pembelajaran Berdasarkan Kurikulum Nasional.* Bandung: PT. Remaja Rosdakarya.

Gierl, Mark J. et al. 2017. Developing, analyzing, and Using Distractor for Multiple-Choice Tests in Educational: A Comprehensive review. *Review of Educational Research*, 87 (6). (online) (http://rer.aera.net) accessed on July, 31st 2019.

Hartati, N., & Yogi, H. P. S. (2019). Item Analysis for a Better-Quality Test. *English Language in Focus (ELIF),* 2(1), 59–70.

Heaton, J. B. 1988. Writing English Language Tests: A Practical Guide for Teachers of English as a Second or Foreign Language. London: Longman.

Hughes, A. 2003. Testing for language teachers. Cambridge: Cambridge University Press.

Karim, S.A., Sudiro, S., & Sakinah, S. 2021. Utilizing Test Items Analysis to Examine the Level of Difficulty and Discriminating Power a Teacher-Made Test. *EduLite Journal of English Education, Literature, and Culture.* 6 (2), 256-269

Luoma, S. 2004. Assessing speaking. Cambridge: Cambridge University Press

Kellaghan, T., & Greaney, V. 2001. Using Assessment to Improve the Quality of Education. Fundamentals of Educational Planning. Retrieved from https://inee.org/system/files/resources/Using_Assessment_to_Improve_the_Quality_of_Education.pdf on 24 April 2021

Kemp, J.E, Morrison, G.R. & Ross, S.M. 1994. *Designing Effective Instruction.* New York: Macmillan College Publishing Company.

Manalu, D. 2019. An Analysis of Students Reading Final Examination by Using Item Analysis Program On Eleventh Grade of Sma Negeri 8 Medan. *JETAL: Journal of English Teaching & Applied Linguistics*. 1(1), 13 – 19.

Mertler, C.A. *2019. Introduction to Educational Research* . Los Angeles: Sage Publication

Ngalimun. 2018. *Evaluasi dan Penelitian Pembelajaran.* Yogyakarta: Parama Ilmu. https://semnas.unikama.ac.id/ks2b/arsip/2017/berkas/28.pdf on 2 April 2021.

Setyawati, U., Putri, S.A., & Puparini, D. 2018. A Quality Analysis Of English Final Test For Third Grade Student At MTS Darul Ulum Karangpandan Rejoso Pasuruan. *JIES.* 9 (2).

Sudijono, A. 2008. *Pengantar Evaluasi Pendidikan*. Jakarta: PT. Radja Grafindo Persada

Sudijono, A. 2011. Evaluasi Pedidikan. Jakarta; Raja Grafindo Persada

Sudjana, N. 2014. *Penilaian Hasil Proses Balajar Mengajar.* Bandung: PT. Remaja Rosdakarya.

Sulistyo, G.H. 2015*. EFL Learning Assessment at Schools. An Introduction to its Concepts and Principles*. Malang: Universitas Negeri Malang.

Sulistyo, G. 2017. *ICT-Based (Authentic) Assessment in the Context of Language Teaching in the Indonesian (Lower And Upper) Secondary Levels of Education: Potential Areas for Real-World Development*. A Paper presented in Konferensi Nasional Sastra, Bahasa & Budaya (KS2B) Universitas Kanjuruhan Malang, p. 239-254.

Sulistyo, G. 2018. *EFL Learning Assessment At Schools An Introduction to Its Basic Concepts and Principles*. Malang: CV Bintang Sejahtera Abadi

Wahyuningrum, Tri. 2017. *Item Analysis Try Out Test for National Final Examination*. Unpublished. S1 Thesis. Pasuruan: English Educational Study Program STKIP PGRI Pasuruan.

Wells, C.S & Wollack, J.A. 2003. *An Instructor's Guide to Understanding Test Reliability*. Testing & Evaluation Services University of Wisconsin. Retrieved from https://testing.wisc.edu/Reliability.pdf on 2 April 2021.