ISSN 2621-6485

The quality of an English summative test of a public junior High School, Kupang-NTT

Thresia Trivict Semiun [a,1,*], Fransiska Densiana Luruk [b,2]

[a, b] University of Timor, El Tari Km 09, Sasi-Kefamenanu-East Nusa Tenggara, Indonesia
[1]tsemiun@yahoo.co.id*; [2] densianaluruk@gmail.com
* corresponding author

ARTICLE INFO

Article history
Received 02 July, 2020
Revised 30 July, 2020
Accepted 30 August, 2020

Keywords
English summative test
validity
reliability
item analysis

ABSTRACT

This study aimed at examining the quality of an English summative test of grade VII in a public school located in Kupang. Particularly, this study examined content validity, reliability, and conducted item analysis including item validity, item difficulty, item discrimination, and distracter effectiveness. This study was descriptive evaluative research with documentation to collect data. The data was analyzed quantitatively except for content validity, which was done qualitatively. Content validity was analyzed by matching the test items with materials stated in the curriculum. The findings revealed that the English summative test had a high content validity. The reliability was estimated by applying the Kuder-Richardson's formula (K-R20). The result showed that the test was reliable and very good for a classroom test. The item analysis was conducted by using ITEMAN 3.0 and it revealed that the test was mostly constructed by easy items, most of the items could discriminate the students, most distracters were able to perform well, and the most of items were valid.

1. Introduction

The scholars are in harmony to say that validity and reliability are the two important criteria for the quality of language testing. Validity is related to "how well what is assessed corresponds with the behaviour or learning outcomes that should be assessed" (Iliya, 2014). It is to see whether or not a test measures accurately what it is intended to measure (Hughes, 1989). Meanwhile, Haryudin (2015) asserted reliability as "the consistency of the examination scores. Also, it refers to the scope to which the test produces consistent results if different markers mark it." According to Bachman & Palmer (1996), reliability is defined as consistency of test scores. Validity and reliability cannot be separated with assessment. Assessment is a scientific method of the evaluation to acquire feedbacks related to the information of teaching and learning, make teachers and students see the achievements and shortcomings clearly, and improve teaching and learning efficiently (Qu & Zhang, 2013). In evaluating students' achievement, a teacher-made test should bear objective and accurate scores. Of course, developing a good test is not easy to do, especially if teachers do not understand or have limited information related to the procedures or principles of a good test. However, if the test is not good, the result yielded by the test is of course not good too. This can harm students since the result is not objective and unfair, and the students' true competence cannot be reflected. Arikunto (2013) stated that teacher-made tests are useful to determine how good students master the learning materials given in a particular time are, to determine whether a learning objective is already accomplished, and to obtain scores. For these reasons, it is important for teachers to possess the skills of creating valid and reliable test and also in analyzing assessments.

Studies in the area of evaluation had been conducted by scholars (Cang & Wu, 2012; Abedi, 2009; Munoz, et al., 2003) on different focuses on teaching and learning English as a foreign language, particularly in countries of non-native speakers of English. Abedi (2009) had conducted research to evaluate assessments for English Language Learning (ELL) students in Turkey. He focused on evaluating language factors based on the assumption that when instructional materials contained complex linguistic structures, ELL students might face serious difficulties in understanding instruction of the test. The result revealed that such unnecessary linguistic complexity seems to affect the validity and reliability of the assessment that lead to the low quality of ELL outcomes. So, a good language is very important to gain a good quality of a test and certainly to avoid frustration for the students as the test takers. The sample size of the study presumably could affect validity and reliability too. Chang and Wu (2012) investigated the validity and reliability of teacher assessment under a web-based portfolio assessment environment. They reported some results of web-based portofolio teacher assessments i.e. (1) achieved an acceptable level of reliability; showed a strong level of inter-rater reliability and inner-rater reliability regarded as a reliable assessment method; (2) demonstrated an acceptable level of validity; (3) and the portfolio scores were highly consistent with the students' end-of-course examination scores, implying that web-based portfolio teacher assessment was a valid assessment method. Based on the results, they confirmed that the smaller sample size might have negative influences on the analysis results. Similarly, Munoz et al. (2003), in their study had acknowledged that the limited number of samples impacted on the less validity of the test.

In the field of English Language Teaching, especially in the Indonesian context, there were valuable studies investigating tests made by English teachers. Primadani (2013) and Ratnafuri (2011) analyzed an English mid-term test and a final test. Both studies revealed that the quality of the tests were not so good due to the reason that the teachers did not follow the rules in how to develop test items which resulted with low quality of the test. Furwana (2019) analyzed the validity and reliability of teacher-made English Summative test in a vocational high school located in Palopo. The result revealed that the teacher made test had good quality regarding content validity and reliability. Another study conducted by Sugianto (2017) was focused on analyzing an English Summative test for senior high school in Palangkaraya. The result showed that the English Summative test was valid and reliable which was proven statistically. To summarize, the valuable studies presented above reported different results on teacher-made tests because constructing a test also related to the competence of the test developers. This indicated that teachers were lack of conceptual assessment tools or the practical skills to investigate or use tests (Fulcher, 2012). The test developers should have been knowledgeable with the issue of constructing a good test. It is important because students' competence will not be reflected truly if the test cannot function properly.

Different from the previous studies, the present study was conducted in Kupang, the capital city of East Nusa Tenggara Province, where the development of education is still the main concern of the government as the former Minister of Education and Culture Muhadjir Effendy (Seo & Setiawan, 2018) said that education in East Nusa Tenggara was the third lowest nationally, after Papua and West Papua. This study highlighted teachers' role in assessing students' achievement through an English Summative test. It was assumed that the result might reveal different outcomes for the quality of the test. The investigation of quality was narrowed to the quality of an English Summative test constructed by an EFL teacher in a public junior high school in Kupang. The goal was to determine if the English Summative test was a reliable and valid measure of students' achievement. The present study investigated the content validity, reliability, item validity including item difficulty, item discrimination, and distracters effectiveness of the teacher-made English Summative test. By investigating the quality of the test, the EFL teacher would be informed and able to do a self-reflection whether the English Summative test had met good quality, or whether the teacher had created a good language test or vice versa. If the test was valid and reliable, the test was useful and truly reflecting the competence of students. In addition, the result of item validity, item discrimination, item difficulty, and effectiveness of distracters could help the EFL teacher to see items that worked well.

2. Research Method

This research was a descriptive evaluative research to describe and evaluate the quality of an English Summative test constructed by an EFL teacher in a public junior high school located in Kupang. This research used documentation to collect the data such as an English summative test, a blueprint, an English syllabus, and students' answer sheets. The data analyses were separated into several parts. First, content analysis was done for revealing the content validity of the test. Within this research, content analysis was defined as the analysis by matching the content of items or questions in the test with the English syllabus used, and the table specification or the blueprint of the test to examine if each item measured the content or objective of the course or unit being taught. Later the proportion of items that measured an indicator would be calculated into percentage. The following considerations were taken as the content review judgments: (1) how appropriate the items are, (2) how complete the item samples are, (3) and the way the items assess the content (Mindes, 2003). Third, the test reliability was done by applying the Kuder-Richardson's formula (K-R20) to obtain inter-item based reliability value. The result of reliability coefficient was interpreted based on interpretation of Nunnally (1978). Fourth, item validity, item difficulty, item discrimination, and distracter effectiveness were analyzed by the means of ITEMAN 3.0 software. The test items were listed according to their degrees of validity (Arikunto, 2013), difficulty and discrimination (Salwa, 2012). Meanwhile, to reveal the effectiveness of distracters, DiBattista & Kurzama (2011) definition was used. A properly functioning distracter was defined as a distracter that had been chosen by at least 5% of the students. If no student chose the distracter, the distracter could not perform well, and that should be removed.

3. Findings and Discussion

The quality of test could be seen through the validity and reliability of the test. Within the test, the quality of the English Summative test of grade VII was evaluated through content validity, reliability, and item quality concerning item difficulty, item discrimination, item distracter and item validity.

3.1. Content Validity

By relying on the content analysis and review judgment, content validity was analyzed. The finding of content validity of English Summative test of grade VII signified that the test had a high content validity as represented in Table 1.

Table 1 displayed the distribution of the items which were in line with the curriculum. 50 items in the test had relevance to the indicators and/or the basic competence meaning that the test had 100% agreement with the curriculum. The findings of the test signified that it was only constructed to measure reading and writing skills as informed by the teacher. The test missed listening and speaking skills due to practical reasons such as time allotment, administration, and cost.

The findings revealed that the test had high content validity. It had 100% agreement with the curriculum. Thus, the test had been constructed with representative samples of materials measured by proper indicators. The English Summative test of grade VII showed high content validity which meant the test was constructed properly. In order to have high content validity, a test should be able to represent the materials given during teaching and learning process in a settled period (Djiwandono, 2011). The materials used were short functional texts and monolog essays. The reading skill such as reading comprehension was appropriate to be measured by multiple-choice form. On the contrary, the writing skill was not appropriate to be measured through multiple-choice form. Brown (2004) asserted that writing was a productive skill and it was best assessed by the product made by the students. It was hard to define students' writing performance by multiple choice. It would be better if the teacher had another type of test to assess students' writing performance.

Table 1. Content validity of the test

Basic Competence	Indicators	Test Item	%
Reading			
1. Responding to the meaning contained in a short functional written text accurately	Given a shopping list, students determine the communicative goal precisely	1	2%
	Provided a greeting card, students determine the implied information accurately	2, 3	4%
	Presented a greeting card, students determine the implicit information in the text clearly	4,5,6,8	8%
	Given a greeting card, students determine the word meaning (antonym) correctly	7	2%
	Provided an announcement, students determine the general description of the text correctly	9,17	4%
	Presented an announcement, students determine the certain information in the text accurately	10,11,12, 18	8%
	Given an announcement, students determine the word meaning (synonym) correctly	13	2%
	Provided a short massage, students determine the certain information in the text clearly	14	2%
	Presented a short massage, students determine the word meaning (synonym) accurately	15	2%
	Given a short massage, students determine the referent precisely	16	2%
2. Responding to the meaning and rhetorical steps of descriptive/procedure texts accurately	Provided a descriptive text, students determine the general picture appropriately	26	2%
	Presented a descriptive text, students determine the communicative goal precisely	19,24	4%
	Given a descriptive text, students determine the certain information in the text correctly	20, 27, 28, 29	8%
	Presented a descriptive text, students determine the main idea of the paragraph correctly	21	2%
	Given a descriptive text, students determine the referent accurately	22, 30	4%
	Provided a descriptive text, students determine the word meaning (synonym) correctly	23, 25	4%
	Presented a procedure text, students determine the communicative goal precisely	31	2%
	Given a procedure text, students determine the certain information in the text accurately	32, 33, 34, 35,36, 37, 38, 39, 40	18%
Writing			
1. Expressing meaning in short functional written text by using a variety of written languages accurately	Provided jumbled words, students can arrange these words into an announcement	41	2%
	Given jumbled words, students can arrange these words into a greeting	42	2%
2. Expressing the meaning and rhetorical steps in a short essay (descriptive/procedure text) by using a variety of written languages accurately	Presented a few sentences, students can arrange these sentences into a coherent descriptive text	43	2%
	Given a few sentences, students can arrange these sentences into a coherent descriptive text	44	2%
	Given an incomplete descriptive text, students can complete the text with the correct nouns and verbs.	45, 46, 47	6%
	Provided an incomplete procedure text, students can complete the text with the appropriate verbs.	48, 49, 50	6%

The result of this present study was in harmony with the result reported by Widowati (2011), Husna (2012), Haryudin (2015), Fathoni (2017), Nugrahanto, et al. (2018) and Furwana (2019).

Thus, teacher-made tests had evidence indicating the right selection of samples in course materials to reveal high content validity. According to Rudner & Schafer (2002), teacher-made tests had the advantage of being directly related to the content already taught in the classroom. The content of tests would be based directly on a detailed course syllabus, books, and other materials used in the classroom. However, in contrast to the result of the present study, Ratnafuri's study (2011) reported moderate content validity of the English Final test, Sugianto (2011) asserted 46% content validity of the English Formative test, Wulandari (2014) stated that the English Summative Test was 51% valid in content, and Setiyana's research (2016) revealed that the validity of the English Summative test was not good since the percentage in content validity was below 73%. A teacher-made test could contain high content validity. Yet, if the content validity was low or moderate then it was presumably related to the competence and/or experience of the test developers in constructing the tests.

3.2. Reliability

The reliability of the test was assessed by evaluating the internal consistency of the test. Based on the inter-item based reliability analysis, it revealed that the reliability coefficient for the test was at .820 so the test was reliable and considered very good for a classroom test. However, there were some items to be revised to maximize the reliability of the tests.

The result was in harmony with the result of Primadani (2013) and Haryudin (2015). The high reliability of the test was due to the number of items which were crucial for test reliability. The teacher-made English summative test of grade VII contained 50 items, so that the test was considered as a long test. According to Griswold (1990) carefully written tests with an adequate number of items usually produce high reliability since they usually provide a representative sample of the behavior being measured. In this regard, Griswold (1990) also said that long tests can make three things to help maintain validity. Firstly, they increase the amount of content that the students must address, ensuring a more accurate picture of student knowledge. Secondly, long tests counteract the effects of faulty items by providing a greater number of better items. Third, long tests reduce the impact of student guessing.

The result of high reliability could be as a result of students who had learned well or because the students remembered the materials given during the instruction. It also could be interpreted that the students had good skills in reading since many students could score high in the test. However, the result of reliability could not be the basis of interpreting students' writing skills.

3.3. Item Analysis

Each conclusion of item difficulty, item discrimination, item distracter, and item validity is provided in Table 2, Table 3, Figure 1, and Table 4.

Table 2. The distribution of classified difficulty index

Range of Difficulty Index	Category	Item
p = 1.00	Very easy	1 Item
.70 < p ≤ 1.00	Easy	33 Items
.30 < p ≤ .70	Moderate	15 Items
.00 < p ≤ .30	Difficult	1 Items
p = .00	Very difficult	0 Item

Item difficulty analysis revealed some results as can be seen in Table 2. First, one item or 2% of the items had the index of difficulty 1.00 (p = 1.00) which meant this item was very easy to be solved by the examinees. So, this item should be removed. Second, 33 items or 66% of the whole items had the index of difficulty .70 < p ≤ 1.00. These items were considered easy and possible to be retained. Third, 15 items or 30% of the whole items had the index of difficulty .30 < p ≤ .70, these items were moderate. Due to this fact, the items were also possible to be retained. Fourth, an item or 2% of the whole items was considered difficult because the index difficulty was .00 < p ≤ .30. However, this item can still be retained.

The findings of item analysis revealed that there were many easy items compared to moderate items. The easy items lead to the interpretation that the items were not changeling for the students hence they could successfully answer the questions. Another interpretation of the findings was the teacher might already give the materials during the instruction. They could answer correctly because they already remembered the answers. There was also the last interpretation, i.e., there were many items that looked easy due to the reason that there were also many good students. As labeled with accreditation A, this school has become a favorite school in Kupang. To be selected as students in this school, the candidates should follow a placement test. They, who were selected, passed the criterion score of the placement test and they had good or high index grade value of elementary national examination. Therefore, most of the students were good students academically.

Table 3. The distribution of classified discrimination index

Range of Discrimination Index	Category	Item
.40 and above	Very good	17 Items
.30 – .39	Good	13 Items
.20 – .29	Sufficient	7 Items
.19 and below	Poor	13 Items

Item discrimination analysis, as shown in Table 3, showed that out of 50 items in the test, there were 17 items or 34% items which were very good in discriminating the performance of the examinees. Meanwhile, 13 items or 26% items could discriminate up and low group of the examinees yet these items were not as good as the previous 17 items. Next, 7 items or 14% items only had sufficient discrimination power, while 13 items or 26% items could not discriminate the performance of the examinees at all. Thus, these sufficient and poor items should be reconsidered.

In the test, there were many items with good discrimination power than poor discrimination power. Although the items were easy, the items still had good discrimination power. Therefore, it could be interpreted that the items looked easy because there were many good students involved in the present study not because the items were below their level of competence. However, it was a need to conduct another study to find out the characteristics of the students involved in the present study.

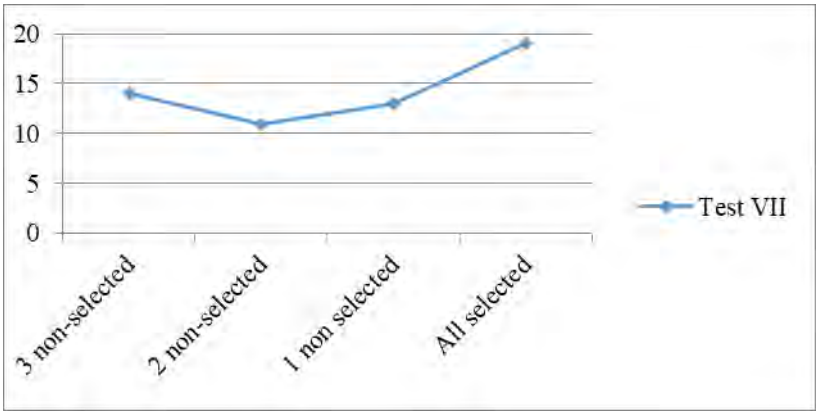

Fig. 1. The distribution of distracters within the test

Figure 1 presents summaries of the findings. The result of distracters analysis asserted that all distracters in item number 10, 11, 12, 29, 37, 38, and 50 were not selected at all. The distracters should be removed because the distracters did not contribute to the questions' discriminatory ability. Next, 22% items (3, 8, 9, 14, 21, 30, 32, 34, 35, 39, and 40) had only one functional distracter and 26% items (1, 2, 5, 6, 16, 22, 28, 33, 36, 43, 44, 48, and 49) had two functional distracters. In these cases, the distracters were reconsidered or replaced with better ones. And, 38% items (4, 7, 13, 15, 17, 18, 19, 20, 23, 24, 25, 26, 27, 31, 41, 42, 45, 46, and 47) had good alternative answers. Hence, all three items can lure the examinees who did not have much information related to the questions.

The analysis showed that there were many effective distracters as well as ineffective distracters within 50 items. The test for grade VII students had 19 items in which all the distracters could work effectively. When all the distracters could function effectively, it could be assumed that the materials tested by the items were new or never be given to students during the instruction.

Table 4. The Distribution of classified validity index

Range of Item Validity	Category	Item
.81 – 1.00	Highly valid	1 item
.61 - .80	Valid	11 items
.41 - .60	Adequately valid	21 items
.21 - .40	Lees valid	11 items
.00 - .20	Poorly valid	6 items

The validity of 50 items (Table 4) showed that out of the 50 items there were 6 items (10, 13, 20, 34, 35, and 38) that should be removed, and 11 items (3, 8, 11, 12, 21, 22, 24, 25, 41, 49, and 50) that should be revised. The rest items could be accepted because they were considered as valid items. Item validity of the test also showed that the number of valid items was greater than the number of invalid items. The valid items had contributed to the reliability of the tests and to maximize the test reliability, invalid items should be removed.

4. Conclusion

The English Summative test to test the achievement of the grade VII students was categorized as a good test, to be specifically presented next. First, in terms of content validity, the test had a high content validity, where 50 items (100%) had an agreement with the curriculum. Second, in terms of reliability criteria, the English Summative test showed reliability coefficient value at .820 indicating that the test was good for a classroom test. Third, in terms of difficulty level, the test was mostly constructed by easy items. The items looked easy presumably because the students were good academically or because the same materials had already been given during teaching and learning instruction. For the discrimination index result, most of the items could discriminate between students who were good and students who were weak. Next, in the case of the item distracter it was concluded that most distracters or alternative answers were able to perform well in the test. Last, the item validity of the test showed that the numbers of valid items were greater than the number of invalid items. The result of the present study indicated that it was important for the teacher to construct an appropriate test. The test used was aimed at measuring the reading and writing skills as intended by the teacher. However, the 10 items used to measure a writing skill might be reliable but it was not a valid test of the writing skill. Therefore, it is a need for the teacher to understand the form of test that appropriately measures the writing skill of the students.

For future researchers who want to conduct research on the same topic, it is suggested to involve experts to validate content validity. The judgments of each item need carefully checked in order to make the relevance of each item with the curriculum more precise. In order to reveal more accurately validity results, it is suggested to examine not only content validity but also face and construct validity to give wider views in regard to the appropriateness of the tests. Last, it will be better if future studies also observe the characteristics of the students. The characteristics of the students will help in interpreting the findings.

References

Abedi, J. (2009). Validity of assessments for English language learning students in a national/international context. *ESE*, *16*, 167-183. https://www.semanticscholar.org/paper/Validity-of-assessments-for-English-Language-in-a-Abedi/0351. Accesed 24 July 2020.

Arikunto, S. (2013). *Dasar-dasar evaluasi pendidikan* (2nd ed.). Jakarta: Paragonatama Jaya.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice. Designing and developing useful language tests*. Oxford, New York: Oxford University Press.

Brown, H. D. (2004). *Language assessment. Principles and classroom practices*. White Plains, NY: Pearson Education, Inc.

Chang, C.-C., & Wu, B.-H. (2012). Is teacher assessment reliable or valid for high school students under a web-based portfolio environment? *Educational Technology & Society*, *15*(4), 265–278. https://www.ds.unipi.gr/et&s/journals/15_4/23.pdf. Accesed 24 July 2020.

DiBattista, D., & Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *The Canadian Journal for the Scholarship of Teaching and Learning*, *2*(2). https://doi.org/10.5206/cjsotl-rcacea.2011.2.4

Djiwandono, S. (2011). *Tes bahasa. Pengangan bagi pengajar bahasa* (2nd ed.). Jakarta: PT. Indeks Jakarta.

Fathoni. (2017). *An analysis on content validity of English summative test items of second grade students at MTsN Kalijambe in the academic year 2015/2016*. State Islamic Institute of Surakarta.

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly*, *9*(2), 113-132. https://doi.org/1080/15434303.2011.642041

Furwana, D. (2019). Validity and reliability of teacher-made English summative test at second grade of vocational high school 2 Palopo. *Journal of Language and Literature, 13*(2), 113-122. https://doi.org/10.15294/lc.v13i2.18967

Griswold, P. A. (1990). Assessing relevance and reliability to improve the quality of teacher-made tests. *NASSP Bulletin*, *76*, 18-24. https://doi.org/10.1177/019263659007452305

Haryudin, A. (2015). Validity and reliability of English summative tests at junior high school in West Bandung. *Jurnal Ilmiah P2P STKIP Siliwangi, 2*(1), 77-90. https://doi.org/10.22460/p2m.v2i1p77-90.167

Hughes, A. (1989). *Testing for language teachers.* Cambridge: Cambridge University Press.

Husna, H. H. (2012). *An analysis of English summative test for 6th grade students in three public elementary schools in Udanawu District, Blitar Regency*. State University of Malang.

Iliya, A. (2014). Formative and summative assessment in educational enterprise. *Journal of Education and Practice, 5*(20), 111-117. https://www.iiste.org/Journals/index.php/JEP/article/view/14252. Accessed 24 July 2020.

Mindes, G. (2003). Assessing young children (2nd ed.). New Jersey: Pearson Education, Inc.

Munoz, A., Alvarez, M., Casals, S., Gaviria S., & Palacio, M. (2003). Validation of an oral assessment tool for classroom use. *Columbian Applied Linguistics Journal*, *5*(5), 139-157. https://doi.org/10.14483/22487085.186

Nugrahanto, A., Winarsih, D., & Farikah. (2018). The content validity of the summative test items of English for the tenth graders of SMA Negeri 1 Magelang in the school year 2015/2016. *Journal of Research on Applied Linguistics Language and Language Teaching, 1*(1), 7-14. https://doi.org/10.31002/jrlt.v1i1.188

Nunnally, J. (1978). Psychometric theory. New York: McGrow-Hill Contemporary.

Primadani, A. E. (2013). *An analysis of a midterm English test of the 7th grade accelerated class at SMPN 3 Malang.* State University of Malang.

Qu, W., & Zhang, C. (2013). The analysis of summative assessment and formative assessment and their roles in college English assessment system. *Journal of Language Teaching and Research*, *4*(2), 335-339. https://doi.org/10.4304/jltr.4.2.335-339

Ratnafuri, W. F. (2011). *An analysis of the teacher-made English test in the final test of the 2nd semester of 2010/2011 of the first year students of SMPN 1 Kauman, Tulungagung*. State University of Malang.

Rudner, L., & W. Schafer. (2002). *What teachers need to know about assessment*. Washington, DC: National Education Association.

Salwa, A. (2012). *The validity, reliability, level of difficulty and appropriateness of curriculum of the English test.* Diponegoro University.

Seo J., & Setiawan, K. (2018). Menteri Muhadjir Effendy: Pendidikan di NTT urutan 2 terbawah. *Tempo.Co*. https://nasional.tempo.co/read/1048094. Accessed 25 July 2020.

Setiyana, R. (2016). Analysis of summative tests for English. *English Education Journal, 7*(4), 433-447. http://www.jurnal.unsyiah.ac.id/EEJ/article/view/5525. Accesed 25 July 2020.

Sugianto, A. (2017) Validity and reliability of English summative test for senior high school. *Indonesian EFL Journal: Journal of ELT, Linguistics, and Literature*, *3*(2), 22-38. http://ejournal.kopertais4.or.id/mataraman/index.php/efi/article/view/3191. Accessed 24 July 2020.

Sugianto, A. (2011) Analysis of validity and reliability of English formative tests. *Journal of English as a Foreign Language*, *1*(2), 87-94. https://doi.org/10.23971/jefl.v1i2.193

Widowati, D. H. (2011). *Item analysis on a teacher-made try-out test of UAN 2010/2011 of junior high schools in Malang.* State University of Malang.

Wulandari A. (2014). *An analysis on the content validity of the summative test items at the even semester of the second grade; a case study of MTs Al-Amanah*. Syarif Hidayahtullah State Islamic University.