FOSTER
Journal of English Language Teaching and Learning

Vol. 1, No. 2, October 2020
e-ISSN: 2723-4126

The Analysis of Junior High School Teacher-Made Tests for the Students in Enrekang

Husni A.
Universitas Muhammadiyah Palopo, Palopo, Indonesia
husni@umpalopo.ac.id

ABSTRACT

The research aimed at finding out information about the preparation of constructing teacher-made tests in Enrekang, the quality of English teacher-made test according to item analysis, and the level cognitive domain of the teacher-made test. The test quality was determined after it was used in school examination test. This research employed survey research using descriptive method. The researcher analyzed the data and then described the research finding quantitatively. The population of this research was the teachers who teach in ninth grade at junior high schools in Enrekang. This research applied simple random sampling technique by taking four different schools as sampel. The results of analysis show preparation that junior high school teachers follow in constructing teacher-made tests in Enrekang is divided into five main parts. In preparing the test, the procedures were considering tests' materials and proportion of each topic, choosing to check the item bank that match to syllabus and indicators, or preparing test specification. In writing test, teachers' procedures were re-writing chosen test item from internet and textbook, re-writing items that was used before and allowing the other teachers to verify it, combining items from item bank and text book, or making new item. While in analyzing a test, the procedures used by the teachers were analyzing and revising test based on its item difficulty, predicting the item difficulty and revising the test, or doing nothing to analyze the test. About the timing in preparing the test, there are three out of five teachers who need only one week to construct multiple choice tests. Besides, there are two out of five teachers who need two weeks to construct multiple choice tests. While the teachers have different ways in providing test based on students' ability. Moreover, the item analysis shows that no test is perfectly good. It was found that almost all tests need to be revised. It was also found that there were only three categories works in all tests based on the cognitive domain of the test namely knowledge, comprehension, and application categories. There was no item belong to analysis, synthesis, and evaluation categories.

Keywords: Teacher-made Test, Test Quality

INTRODUCTION

Teaching and learning process at school has to follow several things to get successful. Before teaching and learning process, the teacher has to consider about designing curriculum, determining teaching and learning objectives, need analysis, and preparing materials (Musbaing, 2020; Nurdin et al., 2019). The next process is the teaching and learning process, including designing lesson plan and teaching practice. The last process is the learning and teaching evaluation. In order to measure students' understanding, teacher needs to conduct a test. Testing can be conducted before, during, and after teaching. Nitko (2001: 5) defines test as an instrument or systematic procedure

FOSTER
Journal of English Language Teaching and Learning

Vol. 1, No. 2, October 2020
e-ISSN: 2723-4126

for observing and describing one or more characteristics of a student using either a numerical scale or a classification scheme. The test is scored by adding the students' point from each question. In this term, students are described using numerical scale.

After conducting a test, it is very important to conduct item and test analysis. These analyses evaluate the quality of the items and of the test as a whole. Such analyses can also be employed to revise and improve both items and the test as a whole(Iksan & Dirham, 2018; Irvy, 2020). Nowadays, teacher-made tests are also used as a complement aspect that determine whether students can pass or not in national examination (Iksan, 2017). The teacher-made tests are considered to have more ability to show students understanding, but then it gets some critical comments, because some researchers claimed that the test is not proper to use. Therefore, we need to conduct item and test analyses to see and prove the quality of the test. Kubiszyn and Borich (2003: 1) stated that tests are only tools, and tools can be appropriately used, unintentionally misused, and unintentionally abused. So it is an important thing to carry out test analysis to see the quality of teacher-made test. In this research, the researcher also tried to analyze the test related to the levels of cognitive domain of the tests. There are six levels of cognitive domain. They are knowledge, comprehension, application, analysis, synthesis, and evaluation (Bloom, 1956; Bloom, Hastings & Madaus, 1971). In this research the tests will be analyzed which one belongs to those levels of cognitive domain. The other problems in trusting teacher-made test is caused by tendency where teachers are increasingly caught up on raising students' scores (Popham 2001: 16).

It is very important for the teacher in particular an English teacher to know how to construct a good test. Constructing a good quality English test instrument, especially for multiple-choice test, is definitely not easy. A trial run must be applied on the freshly designed test before the instrument used. Therefore, the analysis will always be needed to evolve the quality of the English test. To review and revise the English tests instrument mostly are designed and used by the teachers as stated above, therefore, the researcher was interested to conduct a research under the title "*The Analysis of Junior High School English Teacher-Made Tests for the Students in Enrekang*."

METHODS

This research employed survey research using descriptive method. It aimed at giving description about the way teachers prepare the teacher-made multiple choice tests, the quality of the English tests used in school examination test of the second semester in 2019/2020 academic year at several junior high schools in Enrekang, including its validity, reliability, items difficulty, items discrimination, distractors analysis of the tests and level of cognitive domain of the tests. The researcher analyzed the data and then described the research finding quantitatively. The variables of this research were the teachers' way in preparing teacher-made multiple choice tests and test quality. In term of test quality, it consists of some sub-variables. Test quality as whole will be analyzed based on its validity and reliability, item difficulty, discrimination index, and distractors power. Besides that, test item quality was analyzed based on levels of cognitive domain of the tests.The population of this research was the teachers who teach in ninth grade at junior high schools in Enrekang. There were eight junior high schools in Enrekang city. This research applied simple random sampling technique by taking four different schools as sampel.

FOSTER
Journal of English Language Teaching and Learning

Vol. 1, No. 2, October 2020
e-ISSN: 2723-4126

The data that were analyzed in this research was the English tests instrument used in school examination test in the second semester 2019/2020 academic year at four junior high schools in Enrekang city. The data took from the teachers of English for the ninth (IX) grade of each school.There were two instruments that employed in this research namely Document analysis and questionnaire.

Data collected in this research consisted of the following data:

1. Teacher-made test documents
2. Students' answer sheets
3. Teachers' responses in questionnaire

Firstly, the researcher collected and organized the data taken. Then the collected data was analyzed by using the procedures as follows:

1. Teacher's test in preparing teacher-made test.

The teachers' responses from questionnaire analyzed and interpreted to determine and describe teachers' way in preparing teacher made test. The data was tabulated in the table of frequency to identify the modus of preparing teacher-made multiple-choice test.

2. Test quality

Test quality was analyzed using software named ANATES. ANATES is a computer application program that can be used to analyze multiple-choice test. ANATES is easy to learn and to be used. The facilities of this program are scoring system, defining upper and lower group, reliability analysis, validity analysis, distractor analysis, counting item difficulty, and discrimination index.

In determining test validity, coefficient of each test item was compared with r-table to find out the degree of significance. If the correlation is bigger than r-table value, it means that correlation is significant. But if the correlation is smaller than r-table, it is assumed to be insignificant. Furthermore, test is assumed highly reliable if its coefficient is bigger than 0.70. If its coefficient is lower than 0.70, it means the test have low reliability.

In order to get the detail information about the quality of the test, item difficulty of each test was analyzed. The level of item difficulty is too easy, easy, fair, difficult, and too difficult items. A good test should contain easy, fair, and difficult items, not only focused on one level of difficulty. Too easy or too difficult items are not acceptable to be used in a test. The following criterion is used to determine item difficulty.

Item difficulty	Interpretation
0 – 15%	Too difficult
16% - 85%	Acceptable
86% - 100%	Too easy

The next analysis was item discrimination. It was determined the effectiveness of each item to differentiate students based on their knowledge. The maximum item discrimination difference is 100%. It is occurred when all students in upper group answered correctly and all students in lower group answer incorrectly. The detailed criteria that were used to determine the interpretation of discrimination index can be seen in following table.

FOSTER
Journal of English Language Teaching and Learning

Vol. 1, No. 2, October 2020
e-ISSN: 2723-4126

Discrimination index	Interpretation
Negative – 0.29	Improper
0.30 – 1.00	Proper

The other important element in determining the quality of the test items is distractors power. In this stage, performance of each incorrect option was analyzed to find out effective and ineffective distractors. A distractor can be classified as working properly if it is selected by at least 5% students for 3 answer choices and 3% students for 4 answer choices.

3. The level of cognitive domain of test

In order to determine the level of cognitive domain of test, each item of the teacher made multiple choice tests were analyzed by using Bloom's Taxonomy guide to writing question.

RESULTS

Four different tests used in tests were collected and analyzed as sample of teacher-made multiple choice tests. SMP A test consists of 50 items which mainly focused in testing reading test (35 items), vocabulary test (12 items) and grammar test (3 items). 50 items in SMP B test have test components namely reading test (38 items), vocabulary test (10 items), and grammar test (2 items). Furthermore, 50 items of SMP C test consists of reading tests (42 items), grammar test (4 items), and vocabulary test (4 items). While SMP D test has 50 items consists of reading test (41 items), grammar tests (2 items), and vocabulary test (7 items).

The findings for preparation of tests constructions in constructing multiple-choice tests were derived from questionnaire. Therefore the quality of the tests was determined from item analysis. Referring to the problem statements, researcher came up with the research result as follow.

1. The preparation of the teacher in constructing teacher-made tests in Enrekang

a. Preparing the test

The result of questionnaire shows teachers' procedures in writing multiple-choice tests. It is found that teachers have various procedures. There are four out of five teachers re-write test item that have chosen from the item bank, for example teacher from SMP A stated:

> "Saya pilih dari bank soal yang ada hubungannya dengan materi atau sesuai dengan standar kompetensi dan kompetensi dasar"

Moreover there was only one out of five teachers combine tests that were used before and chose new items from internet or textbook and allow their colleagues to check the test items. It helped test writers to identify ambiguous words, irrelevant information, and unrealistic distractors as the teacher from SMP C stated:

> "Kami menyimpan soal-soal yang telah diujikan sbelumnya sebagai bank soal. Seringkali kami mengambil dari sumber lain seperti buku, BSE dsb. Kemudian mendiskusikan dengan rekan sesama guru"

b. Writing the test

The result of questionnaire shows teachers' procedures in writing multiple-choice tests. It is found that teachers have various procedures. There are three out of five teachers combine test item that have chosen from the item bank and textbook test. Before

FOSTER
Journal of English Language Teaching and Learning

Vol. 1, No. 2, October 2020
e-ISSN: 2723-4126

writing the items, the teachers choose the material, write test specification, and then construct test. For example the teacher from SMP B stated:

> "memilih materi esensial, menulis kisi-kisi soal, menulis soal, dan membandingkan dengan soal-soal yang ada pada buku paket dan bank soal"

Furthermore there are two out of five teachers combine test items that used before and new items from internet or textbook and allow their colleagues to check the test items. it helps test writers to identify ambiguous words, irrelevant information, and unrealistic distractors.

c. Analyzing the test

In analyzing the test, none of the teachers conducts complete item analysis to find out their test quality. Most of teachers only consider item difficulty as the most important thing. There are two out of five teachers just predicted the item difficulty and revised test items. As the teacher from SMP C stated:

> "melihat dari indikator. Kesesuain indikator dengan kalimat apa yang sesuai dengan kaidah tata Bahasa dan tidak ambigu"

Moreover, there are three out of five teachers admitted that no item analysis procedure was conducted during the test construction. As the teacher from SMP D stated:

> "tidak pernah menganalisis dan tidak tahu menguji validitas soal. Jujur tidak pernah memantau kualitas tes yang dibuat. Yang terpenting adalah siswa bisa memahami dan bisa memberi jawaban"

d. Timing in preparing test

The result of questionnaire shows how long the teachers prepare the test. Based on the questionnaire, there are three out of five teachers who need only one week to construct multiple choice tests. Besides, there are two out of five teachers who need two weeks to construct multiple choice tests. Actually the teachers realize that their multiple choice tests were not perfectly good, since no summative test was tried out or fully analyzed. Moreover, most of teachers stated that the given time is not enough to construct a good multiple choice test. They need more time to fix the tests.

e. Providing test based on students' ability

Based on the questionnaire that was given to five teachers in five different schools in Enrekang, the teachers have different ways in providing test based on students' ability. SMP A, the teacher provided test based on students' ability by deciding 10 easy items, 25 middle items, and 15 difficult items. Furthermore, teacher in SMP B has no specification in providing test based on students' ability. The teacher only constructs multiple choice tests based on the curriculum without considering students' ability. The teachers in SMP C stated that the proportion of test commonly acceptable items. While the teacher in SMP D stated that all the items are easy. Even though, only few students can answer the items correctly.

2. The quality of teacher-made test in Enrekang based on item analysis

The quality of teacher-made test is based on five categories. The tests' reliability and validity were analyzed to see its test quality as whole. Furthermore, item difficulty,

FOSTER
Journal of English Language Teaching and Learning

Copyright © The Author(s)
Vol. 1, No. 2, October 2020
e-ISSN: 2723-4126

discrimination index, and distractor power were analyzed to describe test item quality. All categories were analyzed quantitatively using ANATES version 4.0.

a. Validity

Validity is determined by the correlation between item score and total score and compared to r-table to see the degree of significance. If an item has high correlation, it means that it has higher validity. In contrary, if the correlation is lower than r table, it means that item validity is low.

Table 1. Validity of the Tests

School	Grade	Category	Item Number	Percentage
SMP A	IX	Valid	23 items	46%
		Not valid	27 items	54%
SMP B	IX	Valid	23 items	46%
		Not valid	27 items	54%
SMP C	IX	Valid	23 items	46%
		Not valid	27 items	54%
SMP D	IX	Valid	16 Items	32%
		Not valid	34 Items	68%
Overall		Valid	85 Items	42.5%
		Not valid	115 Items	57.5%

Based on statistical analysis result, the existing data showed that each teacher-made test used in each school has different degree of validity. It is found that none of the tests are valid entirely. It is also found that the amount of invalid items in each school is different. Based on test analysis, it is found that in SMP A there is 23 valid items (46%) and 27 invalid items (54%). In SMP B, the test consists of 23 valid items (46%) and 27 invalid items (54%). Furthermore, there are 23 valid items (46%) and 27 invalid items (54%) in SMP C. While the test from SMP D has validity with 16 valid items (32%) and 34 invalid items (68%). Overall, there are only 42.5% test items that can be classified as valid and there are 57.5% invalid items.

b. Reliability

Another important characteristic of a measurement procedure is reliability. Reliability is described as the degree of consistency that a test have in measuring students' ability. The reliability was calculated based on students' answer sheets. According to Sudijono (in Jabu, 2008: 124), the acceptable degree for test reliability is 0.70. If the result is greater than 0.70, it means that the test is highly reliable. In contrary, if the result is lower than 0.70, it means that the test is not highly reliable. The following table shows reliability degree of each test.

Table 2. Reliability of the Test

School	Grade	Reliability	Interpretation
SMP A	IX	0.57	Low reliability
SMP B	IX	0.86	High reliability
SMP C	IX	0.62	Low reliability

FOSTER
Journal of English Language Teaching and Learning

Copyright © The Author(s)
Vol. 1, No. 2, October 2020
e-ISSN: 2723-4126

SMP D	IX	0.84	High reliability
Overall mean		0.72	High reliability

Based on the analysis, it is found that there are four test samples have high reliability. They are tests from SMP B and SMP D. Furthermore, SMP A and C test are low reliability. The test reliability is about 0.57 to 0.86. The lowest reliability is found in SMP A that is 0.57. The reliability in SMP D is 0.84, while in SMP C test's reliability is 0.62. Furthermore, SMP B test's reliability is 0.86. The highest reliability is found in SMP B with 0.86. In general, it can be said that only four tests have acceptable reliability and one test is unreliable. Overall, the mean of all tests is 0.72 which can be assumed as high reliability.

c. Item difficulty

Based on the item analysis, 39 (78%) items from SMP A are acceptable and the other 11 (22%) items are unacceptable because they are too easy and too difficult. In SMP B, there are 34 (68%) acceptable items and 16 (32%) too difficult and too easy items. There are 41 (82%) too easy items, 9 (18%) acceptable items in SMP C. Furthermore, there are 37 (74%) acceptable items and 12 (24%) too easy items in SMP D. Overall, 59.5% test item is acceptable in terms of difficulty level. It is also found that there are 30.5% items that interpreted as too easy and unacceptable, while 10% items are considered unacceptable because they are too difficult for the students.

Table 3. Item Difficulty of the Test

School	Grade	Category	Item Number	Percentage
SMP A	IX	Too easy	5 items	10%
		Acceptable	39 items	78%
		Too difficult	6 items	12%
SMP B	IX	Too easy	3 item	6%
		Acceptable	34 items	68%
		Too difficult	13 items	26%
SMP C	IX	Too easy	41 items	82%
		Acceptable	9 items	18%
		Too difficult	0 item	0%
SMP D	IX	Too easy	12items	24%
		Acceptable	37 items	74%
		Too difficult	1 items	2%
Overall		Too easy	61 items	30.5%
		Acceptable	119 items	59.5%
		Too difficult	20 items	10 %

d. Discrimination index

Discrimination index shows the ability of an item to distinguish high achievers and lower achievers. High discrimination index is achieved if items are mostly answered by high achievers. On the contrary, low discrimination index is found if items are answered mostly by lower achievers. Based on the analysis result, the discrimination index is tabulated in the following table.

Table 4. Discrimination Index of the Tests

FOSTER
Journal of English Language Teaching and Learning

Vol. 1, No. 2, October 2020
e-ISSN: 2723-4126

School	Grade	Category	Item Number	Percentage
SMP A	IX	Improper	11 items	22%
		Proper	39 items	78%
SMP B	IX	Improper	11 item	22%
		Proper	39 item	78%
SMP C	IX	Improper	2 items	4%
		Proper	48 item	96%
SMP D	IX	Improper	5 items	10%
		Proper	45 items	90%
Overall		Improper	29 items	14.5%
		Proper	171 items	85.5%

Based on the table, it can be seen that all test have improper items in terms of discrimination index. In SMP A, there are 39 (78%) items that can discriminate the upper group and lower group properly and there are 11 (22%) items that fail in discriminating. In SMP B, there are 39 (78%) items that work properly and 11 (22%) improper items. Furthermore, there are 48 (96%) proper items and 2 (4%) improper items in SMP C. While test result from SMP D shows that there are 5 10(%) proper items and only 45 (90%) improper items. Overall, the amount of items with proper discrimination index is higher than item with improper discrimination index. It is found that there are 14.5% items which have improper discrimination index, and there are 85.5% items with proper discrimination index. In other words, it can be concluded that there are 14.5% items that fail to discriminate the higher achievers and lower achiever and there are 85.5% items can effectively discriminate higher achievers and lower achievers.

e. Distractor analysis

Distractor power is identified by students' response on each item distractors. Distractor is acceptable if it can make the less knowledgeable students to be confused and choose it. Each option of the tests is analyzed to check the effective and not effective items. The findings show that every test has ineffective distractors in it. The following table will summarize distractor power of each test.

Table 5. Distractor Power of the Tests

Schools	Grade	No. of MC items	Effective distractors per item			
			None	One	Two	Three
SMP A	IX	50	4	24	13	9
SMP B	IX	50	16	15	9	10
SMP C	IX	50	8	20	12	10
SMP D	IX	50	15	15	9	11
Overall		200	43	74	43	40
Percentage		100%	21.5%	37%	21.5%	20%

Based on data calculation, it is found that there are 4 items with no effective distractors, 24 items with only one effective distractor, 13 items with two effective distractors, and 9 items with perfect distractors in SMP A test. Test from SMP B has 16 items with no effective distractors at all, 15 items with only one effective distractor, 9 items with two effective distractors, and 10 items with 3 perfect distractors. Moreover, there are 8 items with no effective distractors, 20 items with one effective distractor, 12 items with two effective distractors, and 10 items with three effective distractors in SMP

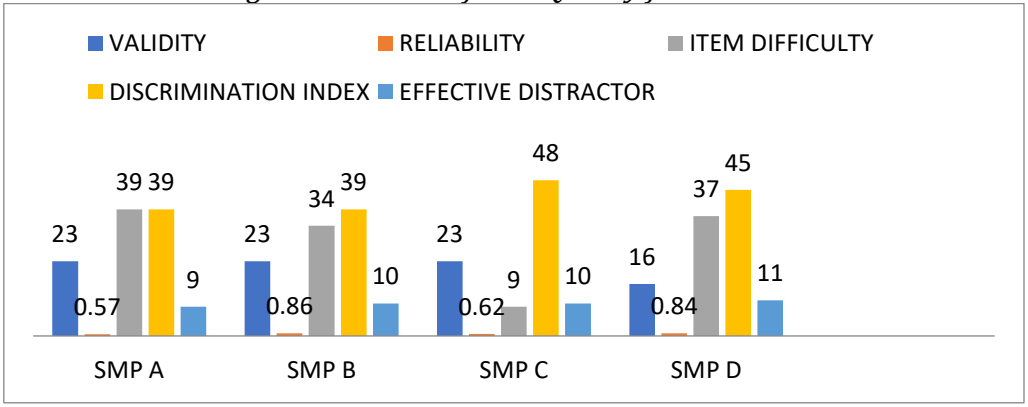
FOSTER
Journal of English Language Teaching and Learning

Vol. 1, No. 2, October 2020
e-ISSN: 2723-4126

C test. While in SMP D test, there are there are 15 items with no effective distractors, 15 items with one effective distractors, 9 items with two effective distractors, and 11 items with perfectly effective distractors . Overall, it can be seen that there are 21.5% items with no effective distractor at all, 37% items with only one effective distractor, 21.5% items with two effective distractors, , and 20% items which all its distractors perfectly work. In general, the amount of ineffective distractors still dominates overall tests.

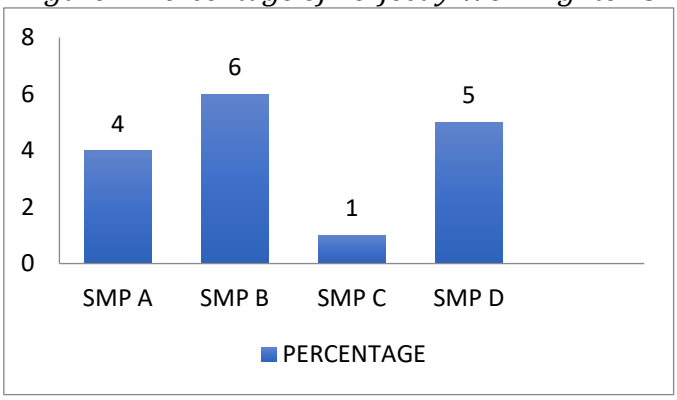
Based on the previous findings, it is found that each test has different quality. The overall findings about tests quality based on validity, reliability, item difficulty, discrimination index, and distractor power are summarized in the following chart.

Figure 1. Results of Test Quality from Each School



After finding each test quality based on item analysis, acceptable items are determined. It is found that there are some test items that can reuse without any revising. In SMP A test, there are only 4 (8%) item out of 50 items that can directly reuse for next test, while there are 6 items (12%) out of 50 items in SMP B that can directly reuse again. In SMP C, there is only 1 (2%) items out of 50 items that perfectly working. Furthermore, SMP D test have 5 (10%) items out of 50 items that perfectly effective. The comparison of perfectly working items from each test can be seen in the following chart.

Figure 2. Percentage of Perfectly Working Items



3. The level of cognitive domain of the teacher-made tests in Enrekang

Four different tests that used in tests were collected and analyzed as sample of teacher-made multiple choice tests. SMP A test consists of 50 items which mainly focused

FOSTER
Journal of English Language Teaching and Learning

Copyright © The Author(s)
Vol. 1, No. 2, October 2020
e-ISSN: 2723-4126

in testing reading test (35 items), vocabulary test (12 items) and grammar test (3 items). And there are 3 items (6%) belong to knowledge, 43 items (86%) belong to comprehension, and 4 items (8%) belong to application term. There is no item belong to analysis, synthesis, and evaluation test. In SMP B there are 50 items have test components namely reading test (38 items), vocabulary test (10 items), and grammar test (2 items). There are 9 items (18%) belong to knowledge, 39 items (78%) belong to comprehension, and there are only 2 items (4%) belong to application. Same in SMP A, there is no item belonging to analysis, synthesis and evaluation test. Furthermore, 50 items of SMP C test consists of reading tests (42 items), grammar tests (4 items), and vocabulary tests (4 items). There are 12 items (24%) belong to knowledge, 35 items (70%) belong to comprehension, and there are only 3 items (6%) belong to application. Same in SMP A and B, there is no item belonging to analysis, synthesis and evaluation test. While SMP D test has 50 items consists of reading test (41 items), grammar tests (2 items), and vocabulary test (7 items). There are 7 items (14%) belong to knowledge, 41 items (82%) belong to comprehension, and there are only 2 items (4%) belong to application. And there is no item belonging to analysis, synthesis and evaluation test.

Table 6. Level of Cognitive Domain of Test

School	Level of Cognitive Domain of Tests					
	know	comp	app	ana	syn	eva
SMP A	3(6%)	43(86%)	4(8%)	0%	0%	0%
SMP B	9(18%)	39(78%)	2(4%)	0%	0%	0%
SMP C	12(24%)	35(70%)	3(6%)	0%	0%	0%
SMP D	7(14%)	41(82%)	2(4%)	0%	0%	0%
Overall	31(15.5%)	158(79%)	11(5.5%)	-	-	-

DISCUSSION

1. The preparation of the teacher in constructing teacher-made tests in Enrekang

Based on teachers' responses on questionnaire, most of the teachers prefer to directly consider which materials or topics they want to use in the tests. Their main reason is practicality. So, they can find test items that related to specific topics. The teachers' only have approximately a week to prepare and constructing their test before handing it over to school administrator. They think this process saves time, so they can finish their test on time. It is an easy way that can be done by the teachers to accomplish the deadline, but we can't see the detailed information about the test. So, we can't prove whether the test items represent the indicators properly or not. From all teachers who respond the questionnaire, there are only 100% respondents who make test specification for their tests (Anggraeni et al., 2020; Sagita et al., 2020). The problem is rising because it is also found that actually the teachers made test specification after the test draft is used. There is a tendency that the test specification is made only to fulfill school administrative requirement. The other three of respondents choose to check item bank that match with the syllabus. Based on the findings, we can assume that most of the teachers didn't provide proper test specification. It should be a great concern since test specification gives the teachers guidance and assistance to construct better test.

Based on the findings in teachers' procedure in writing test, it can be concluded that test is actually not completely made by the teachers. Most of the respondents just adopted test mainly from several textbooks or internet. Their own made tests can be found mostly

FOSTER
Journal of English Language Teaching and Learning

Vol. 1, No. 2, October 2020
e-ISSN: 2723-4126

in grammar section. The rationale for this is they do not have enough time to write their own test. Another excuse is the teachers did not have a lot of choice in deciding how to test their students. The decision about the test form comes from school administrator.

The third process in test construction is test analysis. The analysis process actually should be started by test tryout. Based on teachers' response, none of them tryout their tests before re-use it again. Some of the test items were used before but they never analyzed after using it. Even none of the respondents conduct full item analysis, the questionnaire shows that two respondents feel enough by predicting item difficulty before revising. It can be seen that the main concern on item analysis is only item difficulty. The teachers tend to choose moderate difficulty for their test. Furthermore, three respondents do not take any item analysis for the test. For some reasons, the teachers think that tryout the test draft is not possible to do because they might leak the test. Even so, it does not mean that the teacher should not conduct full item analysis. In constructing summative test, test item can be chosen from the previous test that have been used and analyzed before. So, the teachers should not try out their summative tests and minimized leakage problem.

The fourth is about timing that the teachers need in preparing teacher-made multiple choice tests (Jh & Baderiah, 2020). The result of questionnaire shows how long the teachers prepare the test. Based on the questionnaire, there are three out of five teachers who need only one week to construct multiple choice tests. Besides, there are two out of five teachers who need two weeks to construct multiple choice tests. Actually the teachers realize that their multiple choice tests were not perfectly good, since no summative test was tried out or fully analyzed. Moreover, most of teachers stated that the given time is not enough to construct a good multiple choice test. They need more time to fix the tests.

The last item to be considered in preparing teacher-made multiple choices is providing test based on the students' ability. Based on the questionnaire that given to five teachers in five different schools in Enrekang, the teachers have different way in providing test based on students' ability. SMP A, the teacher has his own way to decide how many easy, middle, and difficult items by using formula 30% easy items, 40% middle items, and 30% difficult items. It means that there are 15 easy items, 20 middle items, and 15 difficult items. SMP B, the teacher provide test based on students' ability by deciding 10 easy items, 25 middle items, and 15 difficult items. The teacher only constructs multiple choice tests based on the curriculum without considering students' ability. The teacher in SMP C state that the proportion of test commonly acceptable items. While the teacher in SMP D stated that all the items are easy. Even though, only few students can answer the items correctly.

2. The quality of teacher-made tests in Enrekang based on item analysis

Teachers sometimes can get the wrong idea about test scores. They may assume that high scores mean good instruction and lower scores mean poor scores. Whereas high score can be acquired from a really easy test which only measuring simple instructional objectives, biased scoring procedures or other factors that influenced the scores, such as cheating or providing unintentional clues to the right answers. Low score can be derived from a really difficult test, trick questions, testing content not covered in class, or other factors that influence the scores, such as grader bias or insufficient time to complete the

FOSTER
Journal of English Language Teaching and Learning

Vol. 1, No. 2, October 2020
e-ISSN: 2723-4126

test (Zimmerman et al, 1990: 17). Therefore, item analysis is conducted to provide evidence that the test is appropriate to reflect students' knowledge.

a. Validity

The validity based on the item correlation and total score found that overall 85% test items have significant validity. It can be caused by some factors, such as well constructed item, word choice, or easy items. Since most of the test items are adopted from textbooks or internet, it is found that all tests mostly have well-constructed items. Based on test samples from those four schools, it is also found that it is consist of easy items and familiar word choices.

b. Reliability

Based on the previous findings, the mean of overall test reliability is classified as highly reliable. However, the degree of reliability coefficient in each test is different. It is influenced by several factors. Students' test scores from SMP A, B, and D are more heterogeneous than students' score from SMP C. There is only low variance of students' score in SMP C. In SMP C, students' scores spread mostly in upper range rather than lower range. In the other words, high scores are dominant in SMP C test. While in the other three tests, students' scores spread in wider range. The other factors, test length and time limit, are not considered to bring a lot of differences in reliability coefficients among tests because test length from all tests is same 50 items and time limit is 90 minutes in all tests.

c. Item difficulty

The ideal test should consist of all easy, fair, and difficult items. Easy item is used to encourage less knowledgeable students in answering the test, while the difficult item is used to set higher parameter in discriminating the more knowledgeable students. Based on item analysis, it is found that there are three schools with easy, fair, and difficult items. They are tests from SMP A, SMP B, and SMP D.

While the test from SMP C almost consist of too easy items. There are 82% tests items belong to too easy items in SMP C. Based on the questionnaire; the teacher from SMP C stated that the important thing in constructing tests is the purposes of teaching learning process from basic competence and standard competence. The teacher decided more too easy items in order to help students answer the test correctly. According to Jabu (2008:42), there are several additional constraints that may need to be imposed on the decision to reject items as too difficult or too easy. We need to include specific content although the items are very easy or very difficult to ensure that the test has face or content validity, provide an easy introduction to overcome psychological apathy on the part of the examinee, shape the test information curve, and consider the availability of items. It can be concluded that the test from SMP C is unacceptable in this research.

The other important thing is item sequence. Based on the analysis, it is found that all tests' sequences are irregular. SMP A test is directly started with fair items and easy items. In SMP B test, difficult items and fair items are ordered interchangeably. Moreover, SMP C test is mostly started with very easy and easy items. While SMP D test is started with fair items, easy and very easy. In reality, it is better to start the test with easy items first then it can be followed by fair and difficult items. This is supported by Soureshjani (2011: 52) who found that items sequence affect foreign language learners' performance. He is also found that students who take easy to difficult test have better performance than

FOSTER
Journal of English Language Teaching and Learning

Vol. 1, No. 2, October 2020
e-ISSN: 2723-4126

students who take difficult to easy items. It is also found that easy to difficult tests may encourage and motivate language testees to take the test with more care and interest. In contrast, difficult to easy tests may cause demotivation, stress, and a set of other negative traits in testees and consequently, they may underperform the test.

d. Discrimination index

Based on data analysis, it is found that percentage of proper discrimination index range from 78% to 96%. All tests have bad items in terms of discrimination power. In general, it can be seen that all test has more proper items than improper items. Therefore, it is also found that improper discrimination index can be seen in all tests. Item difficulty is one factor that affecting degree of discrimination index. Too difficult or too easy items usually have low discrimination index. Other factor that can affect discrimination index is the amount of correct answer in lower group and upper group. If lower group choose more correct answer than the upper group, the discrimination index will decrease. Item with improper discrimination index need to be revised because it fails to discriminate students based on their understanding.

e. Distractor analysis

Distractor is considered working if it is chosen at least by 5% of total testee. Based on the finding, it is found that there are 21.5% items with no effective distractors at all. It is mainly caused by easy item, so students' can easily choose correct answer without being disturb by the distractors. It indicates that distractors fail to do their main task. In order to make better distractors, all ineffective distractors need to be revised.

3. The level of cognitive domain of the teacher-made tests in Enrekang

Based on the finding, most of the items from four examples of teacher-made multiple choices include in comprehension category. It because almost items in four schools consist of reading test. SMP A test consists 35 items reading test out of 50 items and there are 86% items belong to comprehension. In SMP B there are 38 items reading test out of 50 items and there are 78% items belong to comprehension. Furthermore, 50 items of SMP C test consists of reading tests 42 items 70% belong to comprehension. While SMP D test has reading test 41 items and there are 82% belong to comprehension. There is no items belong to analysis, synthesis, and evaluation in four samples of teacher-made multiple choices.

This research finding different with the research finding of Marfuah (2008). In her research stated that based on the distribution of their Cognitive Level proposed by Bloom is not proportional. In the First Semester, the proportion for Knowledge 24%, Comprehension 22%, Application 32%, Analysis 20%, Synthesis 2%, and Evaluation is 0%. In the Second Semester, for knowledge level 48%, Comprehension 16%, Application 14%, Analysis 20%, Synthesis 2%, and Evaluation level 0%.

Constructing multiple choices in analysis category is more complex because it is in the fourth level of student understanding. Analysis needs deep students understanding to answer question in analysis level. According to Bloom, analysis is defined in terms of application and comprehension. Analysis emphasizes the detection of relationships of the parts and of the way they are organized.

Also in synthesis and evaluation levels needs deep students understanding to answer question especially for junior high school students. Bloom explains that synthesis

FOSTER
Journal of English Language Teaching and Learning

Vol. 1, No. 2, October 2020
e-ISSN: 2723-4126

of cognition most clearly calls for creative behavior on the part of the student because it involves newly constructed and oftentimes unique products. While evaluation involves making judgments about the value of knowledge. By definition, evaluation is a form of decision making, done at a very conscious and thoughtful level.

CONCLUSION

Based on the result of data analysis and finding in the previous chapter, the researcher puts forward the following conclusion:

1. There are five main stages that the teachers follow when constructing tests. They are preparing, writing, analyzing the test, considering time, and providing the test based on students' ability. In preparing the test, considering tests' materials and proportion of each topic becomes the teachers' major concern. The other teachers choose to check the item bank that match to syllabus and indicators. While the other combine items that is used before and new items from internet and allow the other teacher check or verify it.
2. The test analysis shows that no tests are perfectly good. It is found that almost all tests need to be revised. Test quality as whole shows 42.5% of all test items are valid. Furthermore, it is found that 2 tests are highly reliable and there are 2 tests have low reliability. Based on the findings, it is found that the proportion of each difficulty level needs to be revised. In terms of discrimination index, it is found that 85.5% items are effectively discriminate higher achievers and lower achievers and there are 14.5% items that fail to discriminate higher and lower achievers. While in distractor analysis, overall it can be seen that there are 21.5% items with no effective distractor at all, 37% items with only one effective distractor, 21.5% items with two effective distractor, and 20% items which all its distractors perfectly work.
3. There are only three categories works in all tests based on the cognitive domain of the test. They are knowledge, comprehension, and application categories. Based on the findings, it is found that the proportion of knowledge is 15.5%, while comprehension is 79%, and there are only 5.5% proportions of application category. There is no item belong to analysis, synthesis, and evaluation categories.

REFERENCES

Alauddin. 2002. *The Analysis of The Teacher Made Multiple Choice English Test for The Students of State SMU in Polewali Sub-District.* Unpublished Thesis. Universitas Negeri Makassar

Anggraeni, W., Wahibah, & Assafari, A. F. (2020). Teachers' Strategies in Teaching Speaking Skills at SMAN 1 Palopo. *FOSTER: Journal of English Language Teaching*, *1*(1), 83–97. https://doi.org/10.24256/foster-jelt.v1i1.9Arikunto, S. 2009. *Dasar-dasar Evaluasi Pendidikan (Edisi Revisi)*. Jakarta: Bumi Aksara.

Bachman, Lyle F. & Palmer, Andrian S. 1996. *Language Testing in Practice: Designing and Developing Language Tests.* New York: Oxford University Press.

Bloom, B.S. 1956. *Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain.* New York: David McKay Co Inc.

FOSTER
Journal of English Language Teaching and Learning

Vol. 1, No. 2, October 2020
***e*-ISSN: 2723-4126**

Gensler, Howard J. 2012. *Valid Objective Test Construction*. St. John's Law Review: Volume 60: Issue 2, Article 2. Retrieved at: http://scholarship.law.stjohns.edu/lawreview/vol60/iss2/2

Hadi, S. 2004. *Statistik* (*jilid 2)*. Yogyakarta: Penerbit Andi.

Heaton, J. B. 1990. *Writing English Language Tests, New Edition.* New York: Longman Inc.

Hughes, Arthur. 1992 & 2003. *Testing for Language Teachers.* 1st & 2nd Edition. Cambridge: Cambridge University Press.

Idaka, I. E., Bassey, S. W., and Ayang, E. 2008. *Test as A Stethoscope: the Need for Adequate Training and Re-Training in Educational Test and Measurement.* Global Journal of Educational Research, Volume 7. Retrieved from http: //www. globaljournalseries.com/ index /index.php /gjer/article/viewFile/110/pdf

Iksan, M. (2017). EMPOWERING BUSINESS GROUP PEANUT TENTENG. *Proceeding International Conference on Natural and Social Science (ICONSS) 2017*, *1*(1).

Iksan, M., & Dirham, D. (2018). The Influence of the Economic Students' Motivations and Language Learning Strategies towards Their English Achievement in STIE Muhammadiyah Palopo. *Ethical Lingua: Journal of Language Teaching and Literature*, *5*(1), 110–121.

Irvy, I. I. (2020). Understanding the Learning Models Design for Indonesian Teacher. *International Journal of Asian Education*, *1*(2), 95–106. https://doi.org/10.46966/ijae.v1i2.40

Izard, John. 2005. *Trial Testing and Item Analysis in Test Construction: Quantitative Research Methods in Educational Planning*. Retrieved from http://www.sacmeq.org/downloads/modules/module7.pdf

Jabu, Baso. 2008. *English Language Testing*. Makassar: Badan Penerbit UNM.

Jh, S., & Baderiah. (2020). Learning Evaluation Management: Improving The Quality of Graduates in State Islamic Institute of Palopo. *International Journal of Asian Education*, *1*(2), 61–72. https://doi.org/10.46966/ijae.v1i2.39

Kiswantani, L. 2010. *An Analysis on Final English Test Validity of the Seventh Year of Bilingual Program of SMPN 2 Jepara Based on School Level - Based Curriculum and the Distribution if Cognitive Level in 2008-2009 Academic Year.* Unpublished Research Paper. Surakarta: School of Teacher Training and Education, Muhammadiyah University of Surakarta. Retrived from http://etd.eprints.ums.ac.id/7178/ on 24 Desember 2010 at 20.32.

Kubiszyn, Tom, and Borich, Gary. 2003. *Educational Testing and Measurement: Classroom Application and Practice, 7th Edition*. United State of America: John Wiley & Sons, Inc.

Kuma. 2011. *Test Construction Skill and Assessment of Factors Affecting Test Analysis and Evaluation Methods: the Case of Three Selected High School in Addis Ababa.* Retrieved from http:// etd.aau.edu.et/ dspace/ bitstream/ 123456789/3702/1/Meseret%20Kuma.pdf.

Mansyur. 2009. *Assesmen Pembelajaran di Sekolah.* Yogyakarta: Multi Pressindo.

FOSTER
Journal of English Language Teaching and Learning

Vol. 1, No. 2, October 2020
***e*-ISSN: 2723-4126**

Magno, Carlo. 2003. *The Profile of Teacher-Made Test Construction of the Professors of University of Perpetual Help Laguna.* UPHL Institutional Journal Volume 1. Retrieved from http:// id.scribd.com/ document _downloads/direct/7790173?extension=pdf&ft=1358914263&lt=1358917873&uahk=WuKrTILhh7GjLfPrkzhWLOxzJNk

Marfuah, R. 2008. *An Analysis of English Test Validity of the First Year of SMAN 1 Purbalingga Based on the School-Based Curriculum and The Distribution of Their Cognitive Level.* Unpublished Research Paper. Surakarta: School of Teacher Training and Education, Muhammadiyah University of Surakarta. Retrieved from http://etd.eprints.ums.ac.id/689/ on 24 Desember 2010 at 20.32.

Mujiyanto. 2007. *Analisis Butir Soal Ulangan Akhir Semester Bidang Studi Ilmu Pengetahuan Alam (IPA) Kelas VIII Semester Gasal Sekolah Menengah Pertama Negeri 1 Sukorejo Kabupaten Kendal Tahun Pelajaran 2006/2007.* Skripsi Unpublished. Semarang: Fakultas Ilmu Pendidikan Universitas Negeri Semarang. Retrieved from http://digilib.unnes.ac.id/gsdl/collect/skripsi/archives/HASH0116/d4cad345.dir/doc on 10 Desember 2010 at 21.31.

Musbaing. (2020). Educational Policy: Understanding Tri Pusat Pendidikan (Education Centers) as Efforts to Reach Educational Objectives. *International Journal of Asian Education*, *1*(2), 53–60. https://doi.org/10.46966/ijae.v1i2.35

Nurdin, K., Muh, H. S., & Muhammad, M. H. (2019). THE IMPLEMENTATION OF INQUIRY-DISCOVERY LEARNING. *IDEAS: Journal on English Language Teaching and Learning, Linguistics and Literature*, *7*(1).

Sagita, R. J., Sahraini, & Syam, A. T. (2020). Designing English Syllabus for Islamic Education Study Program at IAIN Palopo. *FOSTER: Journal of English Language Teaching*, *1*(1), 15–28. https://doi.org/10.24256/foster-jelt.v1i1.4

Nitko, Anthony J. 2001. *Educational Assessment of Student, 3rd edition.* New Jersey: Pearson Education, Inc.

Popham, W. James. 2001. *The Truth about Testing: An Educator Call to Action.* Virginia: Association for Supervision and Curriculum Development.

Rahman, Motiour., and Gautam, Arvind Kumar. 2012. *Testing and Evaluation: A Significant Characteristic of Language Learning and Teaching.* Language in India Journal, Volume 12. Retrieved from http:// www. languageinindia.com/jan2012/motiurtestingevaluationfinal.pdf.

Secolsky, Charles. 1987. *On the Direct Measurement of Face Validity: A Comment on Nevo*. Journal of Educational Measurement Vol.24 No. 1. Retrieved from: http:// staff.neu.edu.tr/ ~cise.cavusoglu/ Documents/ Advaced%20Research%20Methods/Quantitative/Secolsky%20direct% 20 measurement% 20of%20face%20validity.pdf

Sudjiono, Anas. 1996. *Pengantar Evaluasi Pendidikan*. Jakarta: PT. Raja Grafindo Persada.

Sudijono, A. 2009. *Pengantar Evaluasi Pendidikan.* Jakarta: Rajawali Press

FOSTER
Journal of English Language Teaching and Learning

Vol. 1, No. 2, October 2020
***e*-ISSN: 2723-4126**

Suhuri. 2008. *Model Evaluasi Pembelajaran Bahasa Inggris SMA.* Unpublished Dissertation. Yogyakarta: Program Pascasarjana Universitas Negeri Yogyakarta. Retrieved from http://www.damandiri.or.id/detail.php?id=820 on 4 November 2010 at 14.21.

Zhang, Zhicheng., and Burry-Stock, Judith A. 2003. *Classroom Assessment Practices and Teachers' Self-Perceived Assessment Skills*. Published Thesis. Retrieved from http:// www.esf.edu/ assessment/ documents/ assessmentpracticesandtskills.pdf.