The Quality of the Indonesian Language Teacher-Made Tests at Junior School Level

Aris Badara
Halu Oleo University, Kompleks Kendari Permai Blok K3/No. 11, Kota Kendari, Indonesia

Corresponding e-mail: arisbadara71@yahoo.co.id

Abstract: Teachers' ability to compose high quality test is highly important, because with the good quality test, teachers are able to measure precisely the success of the learning process they has done. To be able to know teachers' ability in composing learning outcome test, the research has been done in order to acquire clear information. The measurement of learning outcomes test quality made by teacher can be seen in test item validity, test reliability, test discriminating power, and test difficulty level. The research result shows that the tests made by Indonesia language teacher are: (a) invalid so it cannot become students' learning outcomes measuring instrument, (b) it does not have test validity aspect, (c) it does not fulfil test reliability test. Based on the result, intense training related to test composing technique, test quality analysis, and school's evaluation techniques is necessary.

Keywords: *Indonesian Language subject, teachers-made test, test validity, test reliability*

1. INTRODUCTION

Indonesia Language subject is directed to enhance students' ability to communicate properly, verbally or in writing, and fostering appreciation towards Indonesian literary works. Relation to the learning outcomes, teachers have to follow the development of learning outcomes achieved by students regularly. The information acquired through this research is a feedback towards learning that can become reference in fixing and enhancing learning process in order to get the maximum outcomes.

As learning evaluation implementer, teachers are demanded to have ability in choosing and designing precise evaluation instruments to be used in evaluating learning. One of the common and often used evaluation instruments is test.

The study purpose of teachers-made test is to review every test item in order to get qualified test item before it is used. Besides, the purpose of test item analysis is also to help enhancing test quality through revision or disposing ineffective item, and to understand diagnostic information on every student whether they understand in comprehending the material being taught (Ahiri, 2008: 187).

Teachers' ability to compose high quality test is highly important, because with the qualify test, teachers are able to measure precisely the success of the learning that has been done. Besides, the decision taken based on test result is precise. Teachers-made test quality in schools is not yet known with certainty whether the test used by the teachers in learning evaluation implementation is a qualify test or not.

To be able to know teachers' ability in composing learning outcomes test then a research is conducted about learning outcomes test composing ability, especially test used in school examination. The measurement of teachers-made learning outcomes test can be seen from test item validity, test reliability, test discriminating power and difficulty level test.

Based on above statement, it is necessary to know teachers-made test quality from; (a) validity, (b) reliability, (c) difficulty level, discriminating power and distractors, (d) qualify test requirements

2. THEORETICAL REVIEW

The criteria for good test are as follows.

a. Validity

Validity means the extent to which precision and accuracy of a measuring instrument in performing measuring function. A measuring test or instrument are said to have high validity if the measuring instrument perform its function precisely or provide appropriate measurement result with the intent to do the measurement (Sudijono, 1996: 23).

b. Reliability

Reliability is measurement precision and accuracy in evaluation (Sudjana 1996: 16). Measurement result is credible if measurement to the same subject is conducted a few times and the result acquired is relatively similar.

c. Difficulty Level

Test item difficulty level is student's proportion who answers the test item correctly. The good of difficulty index range from 0.3–0.7. Test items that has difficulty index below or above the criterion (0.3–0.7) can be used if there is consideration of the representation of the measured subject (Ahiri & Anwar, 2011: 229).

d. Discriminating Power

According to Surapranata (2004: 23) one of the quantitative analysis purposes is to determine whether or not a test discriminating the group in an aspect measured in accordance with differences within the group.

e. Distractor Function

A test is said to have distractors if the test is objective. The main purpose of the distractor placement in multiple choice test items is to deceive the students who lack of knowledge in choosing the correct answer (Ahiri & Anwar, 2011: 232).

3. RESEARCH METHOD

3.1 Location

The research is conducted in junior high school in Kolaka District; especially the school have never been experienced test quality analysis.

3.2 Type of Research

The research type is using quantitative paradigm analysis through evaluation approach. Through this research is intended to describe the quality of Indonesia language teacher-made test by analysing test item validity, test reliability, difficulty level, test item difficulty index, test item discriminating power index and effectiveness of distractors

3.3 Research Object

The Object in this research is odd semester final test of Indonesia language subject made by Indonesia language teacher in a few junior high schools in Kolaka District conducted in 2015.

3.4 Population and Sample

The population in this research is all of the answer sheets of the Indonesia language teacher-made test that have been tested on 166 students as population. As sample was taken by 40% so that sample accumulation after being rounded is up to 66 students.

3.5 Data Analysis Technique

To calculate test items validity level used product moment correlation formula for essay test and biserial point correlation formula for objective test. The reliability calculation of objective test using Kuder Richardson formula (KR-20) while reliability of essay test using Alpha Cronbach formula.

4. RESEARCH RESULT

4.1 Objective Test Validity

The validity of the teacher-made test of Indonesia language in this research is seen from four aspects that is empirical validity, difficulty level, discriminating power and distractor. It is described as follows.

a. Empirical Validity of Learning Outcome Test

Empirical validity score of the learning outcome objective test made by Indonesia language teacher shows that from 20 objective test items analyzed, there is 15 items claimed as valid as the following table.

No	Difficulty Index	Test Stems	Categories
1	0,71 - 0,90	1,3,5,6,15,16,18	E
2	0,30 - 0,70	2,4,7,8,9,10,11,12, 13,14,17,19,20	M
3	0,10 - 0,29	0	H

Table 1: Test Item Validity Distribution

No	Analysis	Test Item	Total	Categories
1	Valid	1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 16, 17, 18, 19	15	Usable
2	Invalid	4, 13, 14, 15, 20	5	Unusable
		Total	20	

Based on table 1 noted that from 20 items of objective test made by the teacher shows there are 15 test items indicated valid so the 15 test item has fulfil the qualifying requirement so that learning outcome measured with the 15 test items really describe the students' learning outcome. While the other 5 of 20 test items are unusable to measuring students' learning outcome because they are invalid. It means the 5 test items cannot measure the true students' learning outcomes on Indonesia language subject. Therefore, teacher-made test is not yet fulfil the qualifying requirements because based on the theory, the qualified test is the test that has valid items.

b. Difficulty Level of Objective Test

If a test can be answered correctly by all of the students on every level, it can be said that the test is easy. And vice versa, if the test cannot be answered by all of the students, it means the test is difficult. According to Arikunto (2009: 208) P difficulty index classification is as follows:

If P is less than 0.30 it is classified as hard.
If P (0.30 – 0.70) it is classified as moderate
If P is more than 0.70 it is classified as easy.

The outcome of difficulty level of Indonesia language teacher-made test item can be seen in table 2.

Table 2: Test Item Distribution based on Difficulty Level

E: Easy	7
M: Moderate	13
H: Hard	0

Based on table 2 it is noted that from 20 teacher-made test items, then 7 test items categorized as easy items, the other 13 test items categorized as moderate and there is no test item or 0% categorized as hard.

If it is seen from difficulty level, the test do not fulfil the qualified requirement test, because theoretically the qualified test is the test with difficulty level distributed as easy, moderate and hard with the comparison 30% : 50% : 20% or 20% : 60% : 20%.

The findings of this research shows that learning outcome test cannot measure the clever students ability because the test items only consist of easy and moderate test items and there is no item test categorized as hard items.

c. Discriminating Power of Objective Test Items

The score of discriminating power of teacher-made learning outcome test item calculated by using Ahiri formula (2007: 234) that is $D =$ [illegible]. Theoretically, discriminating power coefficient of teacher-made test item is distributed from coefficient -1.0 to +1.0.

The result shows that from all of the 20 objective test item, discriminating power index distributed from 0.05 to 0.71. From 20 teacher-made objective test items, 15 test items are indicated to have a good discriminating power, it means that the 15 test items are able to distinguish the students who learn and do not learn, or able to distinguish smart students with less smart ones, while the other five test item are unable to distinguish the smart students with the less smart ones

Table 3: Discriminating Power Index Distribution of Test Items

No.	Discriminating Power Index	Test Items	Categories
1	0,40 -	10	VG
2	0,30- 0,39	3	G
3	0,20- 0,29	2	E
4	0,01-0,19	5	NG

Descriptions:
VG: Very Good
G : Good
E : Enough
NG: Not Good

Based on table 3 noted that there are 10 test items or 50 % are categorized as very good because it has discriminating power index up to 0.40 and more. It explains that the 10 test items are able to distinguish the smart students and the less smart ones. So, the 10 test items are acceptable. The analysis outcomes also shows that 3 test items or 15% are categorized as good and acceptable, because it has discriminating power up to 0.30 – 0.39. The analysis outcomes also shows that 2 test items or 10% are categorized as enough, because it is acceptable but need to be revised. The research analysis also shows that 5 test items or 25% are categorized as not good (bad) because it is unable to distinguish the smart students with the less smart ones.

d. Objective Test Item Distractors

Distractors of every test items are considered functional if every distractor are chosen 5% minimum of the amount of test subject. Based on the analysis outcomes of Indonesia language teacher-made objective test distractors shows that from 20 test items, 16 test items or 80% have good distractors because it is chosen by 5% and more of test subject, while the other 4 items have non-functional distractors because it is chosen by 5% and less of test subject. The 4 test items that does not functioning are item 1 distractors b and d only chosen by 1.52% of the test subject; item 4 distractor d only chosen by 3.03% of the test subject; item 5 distractor d only chosen by 3.03% of the test subject; and item 16 distractor c only chosen by 3.03% of the test subject and distractor d only chosen by 4.55% of the test subject. For more details can be seen on table 4.

Table 4: Test Item Distractor Function Distribution

Test Item	Answers	Number of Test Subject	%	Description
1	A*	28	42,42	F
	B	7	10,60	F
	C*	48	72,73	F
	D	1	1,52	NF
4	A	13	19,7	F
	B*	39	59,1	F
	C	24	36,36	F
	D	2	3,03	NF
5	A	4	6,1	F
	B*	57	86,36	F
	C	5	7,58	F
	D	2	3,03	NF
16	A	1	1,52	NF
	B*	60	90,91	F
	C	5	7,57	F
	D	3	4,55	NF

Descriptions:
F : Functioning
NF: Not Functioning

Analysis outcomes of distractor also shows that there are test items with not functioning answer key because it is not chosen by majority or most of the test subject. Those test items are item 5 which the answer key only chosen by 2 test subjects or 3.03%. The distractor b was chosen by 57 or 86.36% of the test subjects. Item 7 which the answer key chosen by 28 test subjects or 42.42% of the test subjects while distractor a chosen by 30 test subjects or 45.45%, item 8 which the answer key chosen by 18 test subjects or 12.12% while distractor a chosen by 48 test subjects or 72.73%, item 9 which the answer key chosen by 37 test subjects or 56.06%, and item 12 which the answer key only chosen by 5 test subjects or 7.58%, distractor c chosen by 47 test subjects or 71.21%. Item 14 which answer key only chosen by 21 test subjects or 31.82%, and item 20 which answer key only chosen by 24 test subjects or 36.36%.

Based on the analysis outcome of the four aspects that are empirical validity, difficulty level, discriminating power and distractor, it is concluded that teacher-made test are not qualified because not all of the items are valid, the difficulty level is still in easy and moderate category. Not all of the test items have ideal discriminating power that is 0.30 and more, and there is still test items do not have functional distractors. There are even test items which the answer key are not functioning because it is not chosen by most of the test subjects.

4.2 Reliability Form Objective Tests

Based on reliability analysis of teacher-made tests; it can be seen that the value of alpha / reliability tests are calculated as a whole on 0.722 point. The results of reliability analysis showed that the Indonesian teacher-made tests are not reliable enough. It demonstrates the test cannot measure precisely and consistently student learning outcomes in Indonesian subject. Indonesian teacher-made tests objective form is considered not eligible enough to measure the student learning outcomes since the tests are not reliable.

4.3 Essay Test Validity

Validity of the Indonesian teacher-made tests in this study viewed from four aspects, namely: (a) the validity of the content, (b) the empirical validity, (c) the level of difficulty, and (d) the power difference.

a. Content Validity of Essay Test

Measuring content validity on this research was done by a reviewing of the test item from Indonesian subject learning experts and the expert evaluation of education. Based on the review from both experts, it obtained information about the content validity of the Indonesian teacher-made tests. The 5 items on the test had reviewed and it can be used as a measurement of student learning outcomes due to meet the standards as a good test.

b. Empirical Validity of Essay Test

The Score of empirical validity from the essay test made by teachers Indonesian, after tested to 66 students, showed that the 5 items of essay test form was declared valid because it has

ISSN: 2502-4124

r count is greater than r table of 0.235 to 0.05 α with n = 66. The result is illustrated in Table 5.

Table 5: Validity Distribution of Essay Test

Item	The results of R-Count	R-Table	Description
1	0,727	0,235	Valid
2	0,305	0,235	Valid
3	0,643	0,235	Valid
4	0,775	0,235	Valid
5	0,813	0,235	Valid

Table 5 shows that 5 items of essay test form which created by the teacher is declared valid. 5th test item has been qualified as a good test due to the test has illustrated the actual student learning outcomes. Therefore, the teacher-made essay test is already qualified as a good test.

c. Difficulty Level of Essay Test

From the analysis of item difficulty level achievement test that made by the teachers showed that 5 test items, variations in the level of difficulty is from 0.390 to 0.926. With details of items that the distress index was 0.71 up 3 point that are items 1, 4, and 5. While grain distress index of 0.30 to 0.70 there are 2 items namely 2 and 3. Criteria difficulty level of test item that ideal according to experts is from 0.30 to 0.70. Results of the analysis of item difficulty level teacher-made tests can be seen in table 6.

Table 6: Test Item Distribution Based on Difficulty Index

No.	Difficulty Level	Item test	Category
1	0,71- 0,90	3	Easy
2	0,30 - 0,70	2	Moderate
3	0,10 - 0,29	0	Hard

Based on Table 6, it is noted that from 5 test item that teacher made, then 3 items categorized as easily, two other items are categorized as test as moderate and not one test item which categorized the difficulty level is hard.

Teachers made the test that viewed from the level of difficulty, then the tests do not qualify of test quality, because it is theoretically that the test that qualify is a test that the level of distress distribution easy, medium and hard, with 30%: 5%: 20% or 20 %: 60%: 20%.

The findings of this study indicate that the results test of Indonesian study cannot measure the ability of students who are good because the test only consists of an easy test items and moderate test items. While difficult test items intended to measure the learning ability of students who are good, as well as to motivate clever students to study harder.

d. Differential Power of Test

The results showed that of the 5 items of essay test, so the test items index difference distribute from 0.308 to 1. From 5 items of the essay test that teachers made, test items stated there are 5 item that a good differential power. Meaning to 5 item such tests can distinguish between students who study with students who are not studying or not smart. For more details, a distribution of differential power index to 5 essay test items can be seen in Table 7.

Table 7: Distribution of Differential Power Index of Essay Test Item

No	Index Differential power	Category
1	0, 40…	Very Good
2	0,30-0,39	Good

Based Table 7, it is known that the 4 items in test or 80% of them are categorized as very good because it has discrimination index of 0.40 to above; and 1 other test or the rest 20% are categorized Good as it has discrimination index of 0.38. Discrimination index by 0.38 point difference can already consider as distinguish between students who are good or smart and the students who are not so smart.

4.4. Reliability Tests

Reliable analysis results of the tests showed that the reliability of teacher-made tests is 0,427. The results of the reliability analysis showed that teacher-made tests are not reliable. In the other words, the test cannot reliably measure precisely and consistently about students' learning outcomes in Bahasa Indonesia's Subject. Teacher-made tests in the form of essay do not qualify as a quality test or a test for use in the measurement of student learning outcomes.

4.5 Discussion

4.5.1. Validity Test

Results of the analysis showed that the empirical validity of 20 test items in objective form, there are 15 items, or 75% declared as invalid. 5 points or 25% declared as invalid. Thus, when viewed from the aspect of the validity of the test empirically, then the teacher-made tests do not meet one of the requirements of quality tests. However in the essay test, From 5 test items tested, all of them expressed empirically valid.

Results of the analysis showed that the distribution of difficulty level from teacher-made test in objective form only consists of two categories; 7 items or 35% are categorized as *easy* and 13 items or 65% are categorized as *medium.* None of them are categorized as *difficult*. Similarly, the essay test, 3 items are categorized as *easy* and the other 2 items are categorized as *medium* items.

The findings of this study indicate that teacher-made tests, is not qualified as quality test due to a quality test is a test that has proportional distribution of difficulty levels in each items. It means, levels of difficulty of test items are *easy, medium* and *difficult.* Anastasi and Urbina (1997: 203) states that if the level of difficulty of 70% (p = 0.70) or above, the test is considered easy, and if the difficulty level of 15% (p = 0.15%) below, the test is considered difficult, the better one is an item that has a difficulty level of 50% (p = 0.50%).

Discrimination analysis results indicated that not all items in teacher-made tests has discrimination index. It is known from 20 test items, there are 13 items have a good discrimination *index* because it has the discrimination index of 0.30 and above, 2 items has *sufficient* discrimination features but needs to be revised as the point is around 0.20 to 0.29 and 5 items have *not good* discrimination index because the index difference is below 0.20. So it should be discarded. All 5 items have a good discrimination index. Moreover, the 5 items has a very *good* discrimination index.

The results showed that the discrimination aspects of the 20 test items made by teachers. Only 13 or 65% were categorized as *good* test items, however, 7 or 35% is categorized as *not good* items because it does not have the ability to distinguish between students who study and students who do not learn. The 7 items must be or revised either partial or total revision.

The results showed that the discrimination aspects of the 20 test items made by teachers. Only 13 or 65% were categorized as *good* test items, however, 7 or 35% is categorized as *not good* items because it does not have the ability to distinguish between students who study and students who do not learn. The 7 items must be or revised either partial or total revision.

Results of the analysis in discrimination level of teacher-made test is based on the consideration of experts, the 20 test items tested, 15 test items *can be received,* which means that 15 test items are categorized having discrimination index, while 5 more items cannot be accepted due to the lack of appropriate discrimination index.

This study also showed the function of distractor in 20 test items, there are four test items have distractors who *do not function properly,* because only been selected below 5% of test participants. The 4 items which distractor does not work properly are; item 1 option B and D which only 1.52% of participants selected them, item number 4 which only 3.03% of participants selected the item in the test, item number 5 option D only been selected 3.03% by participants, and item number 16 option c which only 1.52% of participants selected the items and distractors D that only 4.55% of participants selected during the test.

The study also found that in addition to the dysfunctional answers distractors also include an answer key functions that are *not appropriate.* There are 4 items test which the *right option* does not function well as a key answer and only a few participants selected them by test. Moreover, the distractors are chosen mostly by the participants. These results indicate that there has been the placement of obvious distractor, which show different views and comprehension between the participant and test's maker. The distracters are misunderstood by participants test as the correct answer. While the answer key malfunction is indicated as error in the placement of the answer key. However, after re-examined by the Bahasa Indonesia's teacher, the answer key is true. Therefore, the error is caused by students' who do not master the material tested.

As seen from distractors functioning, teacher-made test from Bahasa Indonesia's teacher does not meet the criteria for a good test or quality test because the test is indicated as qualified test items if the distracters are work properly. According to test evaluation expert, the criteria of good distracters is that if any detractors chosen by 5% of the test participants. According Ahiri (2007: 327-328) that a distractor work properly if it is selected at least by five participants from the top group and preferred by the bottom group. According to Anwar (1998: 151) argues that the alternative answer is a good distractor when the distractors at least have been chosen by 5% of all test participants.

Based on the findings of research and discussion, it can be concluded that the Indonesian teacher-made tests do not meet the standards valid if it is seen from the aspect of content validity, empirical validity, the aspect of difficulty levels, discrimination index and functionality aspects of detractors. As previously stated by experts on criteria test items, a proper test items has not been achieved in test items examined. Therefore, the measurement results using Indonesian teacher-made tests cannot be guaranteed to properly measure students' learning outcomes.

4.5.2. Reliability Tests

Based on the results of reliability analysis of teacher-made tests, either the objective test or essay form, it is known that reliability coefficient is 0.722 for objective test and 0,427 for an essay test. Thus, the teacher-made tests *are not reliable.* Determination of reliability criteria 0.75 value refers to the opinion Naga (1992: 129) who stated that the reliability coefficient of 0.75 is considered reliable, but if calculated using Kuder and Richardson (KR 20). Tests revealed counted reliable if r is greater than 0.75. If r counted less than 0.75 then the test is considered unreliable.

Bahasa Indonesia's Teacher-made tests, both objective form and the form of essays are not qualified when viewed from the aspect of reliability. This result means that the Indonesian teacher-made tests cannot show the consistency of measurement. It also does not show the accuracy of the measurement. A good test or quality test is a test that is reliable, because a reliable test means that a reliable measure of student learning outcomes consistently.

5. CLOSING

Based on the results of the discussion that has been described, it can be concluded as follows:

1. There are some items of Bahasa Indonesia's teacher-made tests in objective form are categorized as invalid so that the content validity of test items cannot be used as a measurement of student learning outcomes.
2. Bahasa Indonesia's teacher-made tests are not yet qualified as good test or qualified if it is seen from the aspect of validity of the test. This is due to not all of the items has good discrimination index and not all distracters are functioning properly.
3. Bahasa Indonesia's teacher-made tests do not meet the test of quality seen from the aspect of reliability tests. It can be seen from the analysis of reliability that only reach 0.722 for objective test and 0,427 for an essay test.
4. Empirically, Bahasa Indonesia's teacher-made tests which have been tested to students cannot be used most of them as a tool to measure student learning outcomes since the items do not meet the criteria as a qualified test.

Based on the above conclusions, in developing teacher's ability to construct a test in order to measure student' learning outcomes, it required more in-depth training related to test preparation techniques, quality analysis test, as well as assessment techniques in schools.

6 REFERENCES

Arikunto, Suharsimi. 2003, *Dasar-Dasar Evaluasi Pendidikan*. Yogyakarta: Bumi Aksara.

Ahiri, Jafar dan Anwar Hafid. 2011, *Evaluasi Pembelajaran Dalam Konteks KTSP*. Bandung: Humaniora.

Anastasi Anne, dan Urbina, 1997, *Psychologycal Testing*. New York: Micmillan Publishing Company.

Azwar Saifuddin. 1998, *Reliabilitas dan Validitas*. Yogyakarta: Pustaka.

Blood Don F. Dan Budd, Wiellin C. 1972, *Educational Measurment and Evaluation*. New York: Haper & Row.

Brown Frederick G. 1983, *Principles of Educational and Psychological Testing*. New York: College Publishing.

Sudijono Anas. 1996, *Pengantar Evaluasi Pendidikan*. Jakarta: Grafindo Persada.

Sudjana. 1996 *Teknik Analisis Regresi dan Korelasi Bagi Para Peneliti*. Bandung: Tarsito.

Sukardi. 2009, *Evaluasi Pendidikan, Prinsip dan Operasionalnya*. Jakarta: Bumi Aksara.

Supranata, Sumarna. 2004, *Panduan Penulisan Tes Tertulis, Implementasi Kurikulum 2004*. Bandung: Remaja Rosdakarya.

Zainul, Asnawi dan Nasoetion, Noehi. *Penilaian Hasil Belajar*. Depdikbud: UT, 1997.