AN ANALYSIS ON VALIDITY AND RELIABILITY OF TEST ITEMS IN FINAL SEMESTER EXAMINATION

Aulia Rahmi, Y. Gatot Sutapa Y., Luwandi Suhartono
English Education Study Program Teacher Training and Education Faculty University of Tanjungpura
Pontianak
Email:drian482@gmail.com

Abstract

The main purpose of this research was to get information about the quality of a teacher-made test for semester final examination from Science and Social majors of eleventh grade of SMA Negeri 5 Pontianak that focused on validity, reliability, item discrimination, item difficulty, and distractor. Case study was used in this study. The test items were teacher-made that consisted of 50 items with 321 test takers. Based on content validity analysis, all the test items matched the test blueprint therefore the test was effective and efficient to measure students' knowledge from both majors. The result was supported by reliability analysis that showed 0.817 and it was categorized as very good for a classroom test. Item difficulty analysis also showed good results that both majors had moderate items as the highest number in the test and the means were also in normal distribution. The finding from distractor analysis was also balance that means still have good effect on the test quality. It is suggested that to discard or revise the distractors which did not function well so the test items are not too difficult or too easy, therefore the revised or new items can be used along with the items which already had good quality in terms of item difficulty, item discrimination, and distraction in the future.

Keywords: **Semester Final Examination, Validity, Reliability, Test, Item Analysis**

INTRODUCTION

Every year, the semester final examination is held twice, in the first and the second semester. It measures students' understanding on the materials that have been delivered in a semester and the form of the test is multiple choice. Teachers should find out about the students' achievement during the semester in order to keep record about how well the students have experienced the learning and by that the teacher can do better teaching preparations for the semester at next year so that teachers can make well-planned assessments and evaluation. However, it is found that test items were not constructed, the test maker did not assessed and evaluate the test items. Test maker just used the previous test items for two years which had been used to grade but not really assessed and evaluated so that the test maker did not know wether the test was good or bad to be administered.

There are two main reasons why doing test assessment is important: to give information about students experiencing the learning and to find out about students' achievement (Harlen, 2007, p. 15). Moreover, government has stated a rule about assessment which is from Minister of Education and Culture of Indonesia number 53 year 2015 article 9, last assessment including semester final assessment and year-end assessment are one of mechanisms of assessment of learning outcomes by educational unit. Assessment is not only about students' learning outcome but also about plans to overcome students' difficulty in facing given tests from school. The results of the tests can be used as indicators to meassure how valid and reliable the test was so that next steps are prepared to solve the drawback to make well-planned assessment. The test maker of Semester Final Examination of the school missed the process of assessment. Assessment of learning itself can provide information about what things worked and did not work during teaching learning process reflected from the students' score. Skipping assessment caused losing important things to be noticed by the test maker to make future learning better and decrease the quality of the test items.

Evaluation is as important as assessment. Evaluation is defined as interpreting the means of data about students' score (Blerkom, 2009, p. 7). The results from students are useful for doing evaluation on approaches of teaching and strategy for next learning (Briggs et al, 2008, p. 29). Good standardised tests, checking systems or school-based evaluations are helpful to find out about students' ability for judgment and using knowledge in recent conditions. Evidence and data collected from assessments and evaluations are used to plan improvement at classroom level by collecting evidence about students' understanding, and changing teaching into more suitable way to identification of learning needs (Center for Educational research and innovation, 2005, p. 33). Besides results, the process during evaluation is also necessary because the results of evaluation can base better improvement but again, the test maker did not only skip the assessment part but also evaluation.

Test is commonly used as a way to assess and interpret students' score for evaluation. A test is a tool to measure and get numeric description of learning with standards (Haladyna, 2004, p. 4). Test is used to measure certain intended learning topics, grade from the students' answer in the test. Briggs et al (2008, p. 32) stated that there are some things need to be prepared before administering a test: decide what knowledge will be assessed from what has been taught in class (content validity), how and what students should do for the test, the way of marking, and the form of the test. Well-prepared test helps the process of assessing and evaluating. Based on the interview, the test maker decided what knowledge would be meassured in the test proven by constructing test blueprints. The test maker also did tell the students what and how the test would be along with the material. The students could find out their grade by counting the correct items the multiply by two since the test was in multiple-choice form.

Test has types and formats. Brown (2003, p. 43-48) stated that there are five types of test: language aptitude test to measure capacity or general ability in learning a foreign language, proficiency test to test competence in a languge which consisted of standardized multiple choice items of skills (grammar, reading, vocab, and aural comprehension), placement test to place a student in a certain level of a language curriculum or school, diagnostic test to diagnose specified aspects of language, and achievement test which is limited to particular material addressed in a curriculum within a particular time frame and offered after a course has focused on the objectives in questions. Haladyna (2004, p. 46) stated that there are some formats of a test: simple observation which simply can be observed whether knowledge is possessed or performed with correct or incorrect answer, simple observation with a measuring instrument, checklist which provide connected similar simple observations which is based on provided items on the checklist, multiple choice which is commonly used for measuring cognitive ability by providing a test items with a right choice and some wrong choices, and essay which measuring cognitive skill too but by making the students wriring the right answer based on standards. Based on these formats, semester final examination takes the form as multiple choice and the type is achievement test.

Doing assessment and evaluation is a must because they can increase the quality of validity and reliability of the test items in semester final examination of the school. Content validity is one of validity aspects that refers to the quality of the test which can exemplify what the test intends to meassure (Boyle and Fisher, 2007, p. 66). The main thing in doing content validity analysis is examining the test specifications or blueprint for constructing test items to make sure that the test items are usable for measuring what the students know. Obvious test specifications can increase the consistency and the validity of the items so that they more likely match with the aims of the the test (Brown, 2005, p. 226). The test specifications are the connector between the content standards and the assessment and give a framework to identify what is measured in the test and what items the test include.

Besides validity, there is also reliability in a test. KR-20 is a method formula that can be used to establish reliability. Fulcher and Davidson (2007, p. 107) stated that Kuder Richardson 20 (K-R20) is an estimate of all possible split halves for a test made up of independently scored items. Kubyszyn and Borich (2007, p. 322) stated that K-R20 is more difficult to caculate since it requires the percentage of passing each item on the test but more accurate. This study used this formula since it was focused on accuracy.

In getting information about a test quality from asessment and evaluation, it can also be established by analysing item difficulty, item discrimination, and distractor. Item difficulty refers to number of students who choose the precise answer from an item which the amount of spread (easy, medium, difficult category of questions) of test items influences the average number of item difficulty (Blerkom, 2009, p. 127). Item discrimination is the ability of an item to differentiate among students on how well they have prepared for the test. Distractor is other provided choices which are wrong in a multiple choice question (Blerkom, 2009, p. 129).

The main point of assessment and evaluation is doing improvement on learning by giving test to judge whether the materials for students to master are achievable. Assessment and evaluation are done by doing analyzing validity and reliability of a test. The results of analysis can provide useful information about what topics are considered difficult by students. Considering carefully at the reason why errors happened and finding out right way of teaching are useful to increase knowledge of what best way the materials should be delivered but the result of observation showed that analyzing test items as a form of assessing and evaluating was not conducted by educator.

RESEARCH METHODOLOGY

In this research, descriptive study was used and the research subject was English Teacher Made-Test of semester final examination for eleventh grade. The study results description included the use of analysis statistically. The researcher used documentary study to establish the validity and reliability as the technique, test blueprints and test items are as the the tools in collecting the data.

This research focused on validity and reliability. The validity was analysed manually that only focused on content validity. The test items were analysed by finding out wether the items and the indicators or blueprints matched. Reliability used KR-20 for the best accuracy, analysed in Master TAP. Meanwhile item difficulty, item discrimination, and distractor were analysed by using Master TAP too. All of the students' answer in the the test written in the sheets were input to Master TAP and it gave the results and the results was analysed.

RESEARCH FINDINGS AND DISCUSSION

Research Findings

The researcher asked for test blueprint to analyze the content validity whether the test blueprints were presented like what had been planned in the test but the test maker could not give it. Based on the information from the test maker, the researcher found that before the day of the test, the test maker had decided materials that wanted to be tested, the test blueprints of the materials that wanted to be measured were prepared before constructing the test. The materials for the test were the same for Science and Social majors. The test items were chosen and taken from previous test. The test that has the exact same items for both majors was given to students from Science and Social class. After the test the test maker gave grade on the students answer sheets and then gave them to the students.

Item difficulty refers to number of students who choose the precise answer from an item. Miller (2008: 131) stated that most norm-referenced test developers recommend a .30 to .70-difficulty range with an average item difficulty of .50 to maintain normal distribution. Master TAP analysis result showed that the mean of item difficulty from united majors (united analysis from Science and Social majors) was 0.564, Science major was 0.600 and Social major was 0.535 that based on the table of criteria level of difficulty and the statement from Miller, the mean of difficulty from both majors were categorized as normal distribution. From 50 test items, the items were categorized as follow:

Table 1. Item Difficulty Analysis Results

	Classification	**Item(s) Number**	**Total**	**Percentage**
United Majors (answer sheets from Science and Social major were analyzed in one analysis at a time)	Difficult	7, 12, 22, 26, 32, 35, 45, 47, and 50.	9 items	18 %
	Moderate	2, 8, 9, 11, 13, 14, 17, 21, 23, 24, 25, 27, 30, 33, 34, 36, 37, 38, 39, 40, 41, 42, 43, and 44.	24 items	48 %
	Easy	1, 3, 4, 5, 6, 10, 15, 16, 18, 19, 20, 28, 29, 31, 46, 48, and 49.	17 items	34 %
Science Major	Difficult	7, 13, 22, 32, 35, 45, 47, and 50.	8 items	16%
	Moderate	2, 8, 9, 11, 14, 12, 17, 21, 23, 24, 25, 26, 27, 30, 33, 34, 36, 37, 38, 39, 41, and 42.	22 items	44%
	Easy	1, 3, 4, 5, 6, 10, 15, 16, 18, 19, 20, 28, 29, 31, 40, 43, 44, 46, 48, and 49.	20 items	40%
Social Major	Difficult	7, 12, 22, 26, 32, 35, 38, 42, 45, 47, and 50.	11 items	22%
	Moderate	2, 3, 6, 8, 9, 11, 13, 14, 17, 19, 25, 23, 24, 25, 27, 29, 30, 33, 34, 36, 37, 39, 40, 41, and 43.	25 items	50%
	Easy	1, 4, 5, 10, 15, 16, 18, 20, 28, 31, 44, 46, 48, and 49.	14 items	28%

All the items were categorized based on the index shown in the Master TAP analysis. Every classification used same formula to get the percentage which was the number of items were divided by 50 (the number of test items) then times 100. The table from Science major part showed that moderate items reached 44% (22 items of the total items). Easy items almost dominated the percentage in this test from this major with 40% (20 items of total items) and the least percentage was from difficult items which only reached 16%. In this table, portion of moderate and easy items almost had the same item number but compared with difficult items, there was a great difference.

The table from Social major part showed that exactly 50% (25 items of 50 items) of the test was moderate items. The second highest percentage was from easy items which reached 28% (14 items). The least percentage was 22% (11 items of 50 items) from difficult items. The number of difficult items and easy items almost the same.

From the results above, the test items in semester final examination contained highest number in moderate index that was the same as the mean of difficulty level from both majors. The spreading of both majors was different based on the tables but there were some items that has same level of difficulty from Science and Social majors. There were 27 items (54 %) of 50 items had same results which further well-preparation of constructing of test items is needed to be done since it can influence the average of the results for the test items and to make sure that the test items are well-constructed to measure knowledge of the students.

Item discrimination is the ability of an item to differentiate among students on how well they have prepared for the test. According to Kubyszn and Borich (2006), there is no estimation of good discrimination index but some experts insist that discrimination index should be at least 0. 30, while others believe that as long as the discrimination index has a positive value the item's discrimination is adequate. The analysis from Master TAP showed that the mean of discrimination index from Science major was 0. 238 and Social major was 0. 329 which means that for both majors, their means of discrimination index were categorized as good. The analysis results of item discrimination was shown as follows:

Table 2. Item Discrimination Analysis Results

	Classification	Item(s) Number	Total	Percentage
United Majors (answer sheets from Science and Social major were analysed in one analysis at a time)	Bad	2, 7, 13, 20, 22, 29, 32, 37, 45, 46, 47, 49, and 50.	13 items	26 %
	Good	1, 4, 5, 14, 15, 16, 17, 18, 26, 31, 33, 34, 35, 41, 42, 43, and 48.	17 items	34%
	Excellent	3, 6, 8, 9, 10, 11, 12, 19, 21, 23, 24, 25, 27, 28, 30, 36, 38, 39, 40, and 44.	20 items	40 %
Science Major	Bad	1, 2, 3, 4, 5, 7, 10, 13, 14, 15, 16, 20, 22, 26, 29, 31, 32, 33, 38, 43, 45, 46, 47, 48, 49, and 50.	26 items	52 %
	Good	6, 8, 11, 17, 18, 19, 25, 28, 35, 37, and 44.	11 items	22 %
	Excellent	9, 12, 21, 23, 24, 27, 30, 34, 36, 39, 40, 41, and 42.	13 items	26 %
Social Major	Bad	2, 7, 12, 15, 22, 32, 35, 37, 42, 45, 46, 47, and 50.	13 items	26 %
	Good	1, 4, 5, 13, 16, 20, 21, 26, 29, 30, 31, 34,	16 items	32 %

		36,41, 43, and 49.		
	Excellent	3, 6, 8, 9, 10, 11, 14, 17, 18, 19, 23, 24, 25, 27, 28, 33, 38, 39, 40, 44, and 48.	21 items	42 %

The table from Science major showed that most of the items had bad discrimination power that was 52% (26 items of 50 items). Items that had excellent discrimination power reached 26% (13 items of 50 items). The least percentage was from good discrimination power with the number of percentage was 22% (11 items of 50 items). Bad discrimination power dominated in this test while discrimination power of good and excellent was almost the same.

From Social major, the results showed that the most percentage was from excellent discrimination power that reached 42% (21 items of 50 items). Good discrimination power was 32% in the test (16 items of 50 items) and bad discrimination power had the least percentage that was 26% (13 items of 50 items). Excellent classification is the dominant discrimination power in this test.

It could be seen that there were some items that had same classification for both majors. From the results, researcher found that the most of the test items had bad classification for Science class but for Social class, most of the items were classified as excellent. Another finding was that for both majors, there were 14 items that had same discrimination index but there was no same items for good discrimination index. Therefore, overall, the test items could discriminate between upper and lower group in Social class but it could not discriminate well because of the highest number of items were in bad index in Science class.

In multiple-choice test, well-prepared test takers are supposed to have ability to determine which option is the correct answer and which ones are clearly wrong but less-prepared students see the distractors as potential correct answers. In this analysis, the test consisted of five options which one option as a key answer and other four options as distractors. According to Blerkom (2009), a distractor must have a higher proportion of the less-prepared than the well-prepared students and it has close relation with item discrimination index. When distractor is chosen to be the right answer, then the answer certainly is the wrong answer. From 50 test items, the result was shown as follows:

Table 3. Distractor Analysis Results

	Distractor Analysis	**Classification**	**Item number**	**Total**	**Percentage**
United majors (answer sheets from Science and Social major were analysed in one analysis at a time)	Good	All distractors functioned well	1, 3, 5, 6, 9, 10, 14, 16, 18, 20, 23, 24, 25, 27, 28, 29, 31, 33, 34, 39, 40, 43, 44, 48, and 49.	25 items	50 %
	Bad	One bad distractor	2, 4, 7, 8, 11, 12, 13, 15, 17, 19, 21, 30, 32, 35, 36, 37, 38, 41, 42, 45, and 46.	21 items	42 %
		Two bad distractors	22, 26, 47, and 50.	4 items	8 %

		Three bad distractors	-	-	-
		Four bad distractors	-	-	-
Science Major	Good	All distractors functioned well	19, 21, 23, 27, and 44.	5 items	10%
	Bad	One bad distractor	1, 6, 10, 18, 24, 28, 30, 31, 33, 34, 39, 40, 43, 46, and 49.	15 items	30%
		Two bad distractors	2, 3, 5, 7, 8, 9, 11, 12, 14, 15, 16, 17, 22, 25, 26, 29, 32, 35, 36, 37, 38, 41, 42, 45, and 50.	25 items	50%
		Three bad distractors	4, 20, 47, and 48.	4 items	8%
		Four bad distractors	13	1 item	2%
Social major	Good	All distractors functioned well	1, 3, 5, 6, 9, 10, 15, 16, 18, 20, 24, 25, 27, 28, 31, 33, 39, 40, 44, 48, 49, and 50.	22 items	44%
	Bad	One bad distractor	2, 4, 7, 8, 11, 12, 14, 17, 19, 21, 23, 29, 34, 36, 37, 38, 41, 42, 45, 46, and 47.	21 items	42%
		Two bad distractors	13, 22, 26, 30, 35, 43, and 50.	7 items	14%
		Three bad distractors	-	-	-
		Four bad distractors	-	-	-

From Science major, the number of items that had all the distractor functioned well was 5 items (10 % of the items). The rest 45 items of in the test had bad distractors. This major has all categories of bad distractors with the number of most bad distractor was 50% from two bad distractors category. There was a great difference between good and bad distractors from this major.

From Social major, the table showed that the percentage between good and bad distractors had slight difference with 44% of good distractors and 56% bad distractors but still bad distractors had the most number of the items. this major only had two category of bad distractors with one bad distractors category reached 42% and two bad distractors reached 14%. There were no items in three and four bad distractors category.

From the results above, researcher found that 90% of test items from Science major had bad distractors, only 10% of th test items that all worked as good distractors in the test. The result from Social major had better percentage of good distractors than Science major that the percentage of good distractors reached 44% also it had only one and two bad distractors categories. Compared to Science major, it had all that bad distractors category with two bad distractors category had half of the test items. However, these both majors had few same items. 18% of the test items had same category distractor analysis.

K-R20 is an estimate of all possible split halves for a test made up of independently scored items. Good reliability interpretation is at least on 0.70. If the index is lower than 0. 70, the test item is less reliable. The results from Master TAP showed that KR-20 index from Science major was 0. 739 that was interpreted as good for a classroom and from Social major was 0. 840 that was interpreted as very good for a classroom test. For Science major, if the test maker wants to obtain a KR-20 Reliability of .80, the test must be 1.42 times longer, for a total of 71 items of similar quality to those in the test now and if the test maker wants to obtain a KR-20 Reliability of .90, the test must be 3.19 times longer, for a total of 159 items of similar quality to those in the test now. As for Social major, if the test maker wants to obtain a KR-20 Reliability of .80, the test must be 0.76 times as long, for a total of 38 items of similar quality to those in the test now and if the test maker wants to obtain a KR-20 Reliability of .90, the test must be 1.72 times longer, for a total of 86 items of similar quality to those in the test now.

Discussion

Content validity requires test specifications in constructing a test and reliability deals with the consistency of the test in measuring students' knowledge. These two aspectscan not be separated in analyzing a test. Obvious test specifications can increase the consistency and the validity of the items so that they more likely match with the aims of the the test (Brown, 2005, p. 239). Both Science and Social majors showed good reliability from the test items. The test even had better reliability index in Social major than Science major. The test was good for classroom testing because of the test that was used in SMA Negeri 5 Pontianak Academic Year 2016/2017 for semester final examination matched the tests blueprints. This made the test was effective and efficient to measure students' knowledge from both majors. This condition was proven with a study from Handayani (2009) showed that that good test items are ones which are constructed based on Standard and Basic Competence because it truly can measure students' performance.

The results from item difficulty analysis of 50 multiple-choice summative test items for eleventh grade of SMA Negeri 5 Pontianak with 321 test takers in academic year 2016/2017 showed that both majors had moderate items as the highest number in the test and the means were also in normal distribution based on the shown scale that were included in recommended normal range from Miller (2008, p. 133) in the findings. However, the number difference between difficult and easy items varied from both majors for Science major had greater difference than Social major's. Easy items can make students put little effort or even underestimate in answering the item while difficult items can make students are desperate to answer them.

Although the result of items discrimination from both majors were opposite, the means were at normal distribution which was in line with range stated by Kubyszn and Borich (2007, p. 207) in findings and it meant that the items had good power to discriminate between lower and upper group and the items were effective as classroom testing.

As for the distractors, some of them failed to differentiate lower and upper group since there were upper students chose distractors. These kind of distractors were tricky to be put in the test and revision to increase the effectivity is necessary.

Overall, the test was good for classroom testing. All of the test items matched the test blueprint. Majority of the items were in moderate category and normal distribution. Most of the items had all their distractors worked well but the fact that this test was not assessed and evaluated did contributed in decreasing the quality of the test. From validity and reliability aspects did not have problem since it was valid based on content validity analysis and reliable based on calculation results that showed very good category. There are items which need follow ups whether to be eliminated or revised in terms of item difficulty, item discrimination, and distractor. The analysis results of these three aspects were taken from distractors and alternatives so that the errors of one aspect may affect the quality of other aspects.

CONCLUSION AND SUGGESTION

Conclusion

The test was good for classroom testing for Science and Social majors for eleventh grade of SMA Negeri 5 Pontianak. The test was valid and reliable. Most of the test items had good quality in terms of item difficulty, item discrimination and distractor based on analysis result.

Although the test maker missed assessment and evaluation, the test maker made a good design of a test proven by the analysis results and it must be preserved. Many items can be kept in items bank and used in the future. However, the fact that the results of the findings showed many good items had good quality should not make the test maker miss assessment and evaluation in the future.

Suggestion

Based on the conclusion above, the researcher would share some suggestions for the test maker. It is suggested that to make a test blueprint along with standard and basic competence in constructing a test to see clear measurement of students' knowledge based on material that has been taught. Doing analysis to the test items can help to determine what kind of test items that will work well for Science and Social major. Constructing the test early before the due of semester final examination makes the test maker much time to construct a valid and reliable tes. It also suggested to revise or discard error items and keep the good ones with good results of item analysis. The teacher should consider about reducing her teaching hours so that she has enough time to do assessment and evaluation since it is a must for a teacher.

Assessing and evaluating may consume alot of time and lots of work but since there is a regulation stated that they are a must, there is no reason for test maker to skip doing assessment and evaluation because doing analysis can increase the quality of the test items themselves.

BIBLIOGRAPHY

Blerkom, M. L. (2009). **Measurement and statistics for teachers**. New York: Routledge.

Boyle, J., & Fisher, S. (2007). **Educational testing**. Oxford, UK: Blackwell Publishing.

Briggs, M., Woodfield, A., Martin, C., &Swatton, P. (2008). **Assessment forlearning and teaching in primary schools (2nd ed.)**. Exeter, Great Britain: Learning Matters Ltd.

Brown, H. D. (2003). **Language assessment: principles and classroom practices. Longman.**

Brown, J. D. (2005). **Testing in language programs: a comprehensive guide to english language assessment**. New York: McGraw-Hill Companies, Inc.

Center for Educational Research and Innovation. (2005). **Formative assessment: improving learning in secondary classroom**. Organization for Economic Co-Operation and Development.

Fulcher, G. & Davidson, F. (2007). **Language testing and assessment: an advanced resource book**. New York: Routledge.

Haladyna, T. M. (2004). **Validating and developing multiple-choice test items (3rd ed.)**. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Handayani, H. (2009). **An analysis of english national final exam (UAN) for junior high school viewed from school based curriculum**. Semarang: Diponegoro University

Harlen, W. (2007). **Assessment of learning**. London, Great Britain: Sage Publications Pvt Ltd.

Kubiszyn, T., &Borich, G. (2006). **Educational testing and measurement: classroom application and practice (8th ed.)**. New York: Wiley.