AN ANALYSIS ON VALIDITY AND RELIABILITY OF TEST ITEMS IN PRE-NATIONAL EXAMINATION TEST SMPN 14 PONTIANAK

Hanny Pradana, Gatot Sutapa, Luwandi Suhartono
Sarjana Degree of English Language Education, Teacher Training and Education Faculty,
Tanjungpura University
Emai : hannypradanaa@gmail.com

Abstract

The main objective of this research is to find out whether the test items made by the teachers of SMPN 14 Pontianak were already valid and reliable especially for pre-national examination test. This test was analyzed by case study research with documentary study. The test items used as the sample was English test in pre-national examination test designed by teachers of SMPN 14 Pontianak. The study analyzed pre-national examination test which consist of 50 test items with a total participant of 344 students. The findings were divided into three parts: the appropriateness of test items with the principle of test construction, the validity and the reliability. For the appropriateness of the test items with the principle, the test did not fulfill the principle because the test items were constructed without test indicator. In content validity, the test items also did not show the specific learning purpose of the material being tested. However, the test score reliability calculated using by K-R20 was 0.798, which means the test was categorized as reliable and good to use as diagnostic test. Finally, the researcher recommended the teachers to revise the test items especially errors on item difficulty, item discrimination and distractor analysis.

Keywords: Pre-National Examination Test, Test Item Analysis, Validity, Reliability, Junior High School, Pontianak Regency

INTRODUCTION

Assessment process is used to evaluate students' achievement which can be given in the form of test. According to Cabinet Minister Regulation of Education and Culture No 53 Year 2015 about Learning Achievement Assessment by Teacher and Education Unit for Primary and Middle Education, the teacher's role was making an assessment to assess students' achievement which appropriates with the competence that being assess. Although it is already stated clearly in the Cabinet Minister Regulation, many teachers still reused the test item from the previous test which can make the teacher lose their creativity in making the test. Assessment is a broader term than tests and encompasses the general process of collecting, synthesizing and interpreting formal and informal measurement data (Miller, 2008).

The Pre-National Examination was the test which was held before National Examination. To find out the difficult material students' face to pass National Examination, assessment and evaluation in Pre-National Examination test needed to be done. As a teacher, knowing about the students' achievement through the tests was a must especially it related to National Examination. It is about how good the items made by the teacher based on the validity and the reliability analysis after the test was given. The teacher also should follow the principle of test construction to make sure the test items exactly determine the learning purpose. It was consisting of three

aspects: material, construction and language (Depdiknas, 2011). In material, the test items must be made exactly suitable with the indicator of competence attainment and each test item must have only one right answer or the most right answer. According to Clay (2001), the standard for this principle is the test questions will permit students to demonstrate their knowledge of challenging and important subject matter. And also according to Depdiknas (2011) the teachers should make the indicator which referring to Basic Competence and pay attention on context or material chosen. In construction, the main question must be formulated clearly. The test items must not give hint to the right answer and do not have ambiguous meaning. According to Miller (2008), who stated that multiple choice stems should be free from irrelevant information. In language, the teachers must write the test items with communicative language to avoid misunderstanding. Depdiknas (2011), stated that test items must use communicative language and not repeat the word or phrase which is not part of united concept. It is also supported by Clay (2001), who stated that the standard of language appropriateness of test items is the language must be clear for the assessment tasks and the students.

Validity in test items refers to reliability of test items in measuring students' ability. Validity refers to the degree to which assessment scores can be interpreted as a meaningful indicator of the construct of interest (Young et al, 2013). It consists of several results: content validity, item difficulty, item discrimination and distractor analysis. Content validity is defined as any attempt to show that the content of the test is a representative sample from the domain that is to be tested (Fulcher and Davidson, 2007). To analyze content validity, the area is about what is measured by the test and make judgments about content validity. Content validity is the most common validation that the teachers use to ascertain if a test provides an accurate assessment of instructional objectives (Miller 2008). Item difficulty refers to items with one correct alternative worth a single point; the item difficulty is simply the percentage of students who answer an item correctly (Scorepak, 2015). Item difficulty is relevant for determining whether students have learned the concept being tested. It also plays an important role in the ability of an item to discriminate between students who know the tested material and those who do not. Item discrimination refers to the ability of an item to differentiate among students on the basis of how well they know the material being tested. In item discrimination, it is related to discriminate between strong and weak student in other words we can say that the upper group and the lower group. Strong students or upper group here mean the test takers who have many correct answers in their test. The weak students or lower group is the test takers that have less correct answers in their test. The last analysis for validity in test item is the distractors. In multiple-choice testing, the intended correct option is called the 'key' and each incorrect option is called a 'distractor' (Fulcher and Davidson, 2007).

Reliability in test refers to test items which have consistent result in measuring students' achievement. The standard of reliability is answers to test questions will be consistently trusted to represent what students know (Clay, 2001). It was consisting of several results: K-R21, SEm, and source of error. Kuder Richardson 20 (K-R20) is an estimate of all possible split halves for a test made up of independently scored items (Fulcher and Davidson, 2007). According to Kubiszyn and Borich (2006) K-R20 is the most

accurate formula to measure reliability but difficult to calculate. It is also supported by Brown (2005), if the accuracy is the main concern, K-R20 formula clearly should be used. Standard error of measurement (SEm) is a quantitative expression of the magnitude of error in a test score based on the test reliability (Miller 2008). SEm is used to identify confidence limits around a students' test score. Tests with a large SEm should be carefully reviewed and revised. According to Brown (2005), the function of SEm is to determine a band around the students' score. That's why SEm can provide meaningful statements about the accuracy of test scores obtained by students. The last analysis in reliability was source of error. In source of error the researcher did 4 kinds of analysis, they were: error within test taker which can be called intra-individual error (Kubiszyn and Borich, 2007); error within the test happens when the test are poorly designed obtaining score lower than a true score and the poorly written test repete with clues that also might obtain score higher than a true score (Kubiszyn and Borich, 2007); error within test administration which only focuses on physical comfort, include: room temperature, lighting, noise and seating arrangement are all potential source of error for the students (Kubiszyn and Borich, 2007); and error in scoring which introducing possibilities of error (Kubiszyn and Borich, 2007).

By conducting this research, the researcher hopes that the teacher could realize the importance of assessment and evaluation in learning to find out students' learning achievement. Based on that point of view, this research concerned on: the appropriateness between the test items and the principle of test construction; the result of validity analysis; and the result of reliability analysis in pre-national examination test which consist of 50 items with total participants 344 students.

METHOD

In conducting this research, several methods needed to be applied. They were case study method and documentary study. The case study design is based upon the assumption that the case being studied is a typical case of a certain type and therefore a single case can provide insight into the events and situations prevalent in a group from where the case has been drawn (Kumar, 2011). Documentary analysis involves the study of existing documents, either to understand their substantive content or to illuminate deeper meanings which may be revealed by their style and coverage (Ritchie and Lewis, 2003).

The research subject in this research is the three English teachers in SMPN 14 Pontianak who made the test items of Pre-National Examination Test. The teachers also the people who teach in grade IX in SMPN 14 Pontianak. To get data in this research, the researcher took the indicator of competence attainment and the item test that used for Pre-National Examination Test. After the Pre-National Examination Test already conducted, the researcher took the students answer sheet. The researcher then input the data into the software program, Master Tap.

For validity in test items, the researcher focused on several aspects, they were: content validity, item difficulty, item discrimination and distractor analysis. To analyze content validity, the researcher checked manually the test items used for Pre-National Examination Test whether the test items were suitable or not with the indicator of competence attainment. The researcher checked the test items one by one based on the indicator of competence attainment, the material and the standard competence.

To analyzed item difficulty, the researcher used the software program Master TAP. Firstly the researcher input the data from students answer sheet. After that the researcher concluded the result of item difficulty analysis from Master TAP.

Item difficulty calculated by this formula (Brown, 2005):

$$IF = \frac{\text{N correct}}{\text{N total}} \quad \text{..... (1)}$$

Where

IF	: item facility (item difficulty)
N correct	: number of students answering correctly
N total	: total numbers of students taking the test

The most norm-referenced test developers recommend a .30 to .70 difficulty range with an average item difficulty of .50 to maintain a normal distribution (Miller, 2008). If the proportion of correct answer is less than .30 it is considered as *too difficult,* but if the proportion of correct answer is more than .70 it is consider as *too easy*. Beside analyzing the item difficulty, the researcher also analyze the item discrimination.

Item discrimination calculated with this formula (Brown, 2005):

ID = IF upper - IF lower …….. (2)

Where

ID	: Item discrimination
IF upper	: Item facility (item difficulty) for the upper group on the whole test
IF lower	: Item facility (item difficulty) for the lower group on the whole test

There is no single answer about good discrimination index, but some experts insist that item discrimination should be at least .30, while others believe that as long as item discrimination in positive value, the ability is adequate (Kubiszyn and Borich, 2007). In this stage the researcher also analyzed how effective the distractors affected the key answer. The researcher analyzed it manually and also saw the result on Master TAP. The bad distractor was the distractor chosen by more students in the upper group rather than in the lower group (Kubiszyn and Borich, 2007).

For reliability in test items, the researcher focused on several aspects, they were: Kuder Richardson 20, standard error of measurement, and source of error.

Kuder Richardson 20 (K-R20) Formula (Brown, 2005):

$$\text{K-R20} = \frac{k}{k-1}\left(1 - \frac{\Sigma Si^2}{St^2}\right) \quad \text{..... (3)}$$

Where

K-R20 : Kurder-Richardson 20
k : number of items
Si2 : item variance
St : test score variance

The good test reliability start from index 0.70 until 0.90 and above. The researcher conclude that the interpretation as reliable test. For index 0.50 until less than 0.70 is marked as less reliable test. Another analysis for reliability was SEm.

The formula for calculating SEm according to Miller (2008) is:

$$SEm = SD\sqrt{1 - r} \quad \text{..... (4)}$$

Where

SEm : Standard error of measurement
SD : Standard deviation
r : Reliability coefficient for test

Before the researcher analyzed the SEm, firstly the researcher would draw the skewed of score distribution whether it is positive skewed or negative skewed. The last aspect to analyze in reliability was the source of error. There were 4 observation sheets that the researcher filled to investigate the source of error such as: Observation Sheet 1 (Error within Test Taker), Observation Sheet 2 (Error within The Test), Observation Sheet 3 (Error within Test Administration-Physical Comfort) and Observation Sheet 4 (Error in Scoring) (Kubiszyn and Borich, 2007).

RESULTS AND DISCUSSIONS

Results

The principles consist of material, construction and language (Depdiknas, 2011). For material aspect, the researcher found that all of the test items were not suitable with the indicator of competence attainment. The test items were just made without indicator of competence attainment. Another result in material aspect the researcher found was all of the test items had only one right answer. There was no miskeying that made the test items had the key answer more than one. For construction aspect, it was divided into six parts of analysis. First analysis, the researcher found that all of the test items were formulated clearly. Second analysis, the researcher found that all of the test items were free from ambiguous meaning. Third analysis, the researcher found that 78 % of test items had the same length of options. Forth analysis, the researcher found that 94 % of test items were free from incorrect words. Fifth analysis, the researcher found that only 16 % of test items were tricky test items. The last analysis, the researcher found that there were 18 % of test items less suitable for students in Junior High School because most of the test items contain less familiar vocabulary for students. For language aspect, it was divided into three parts of analysis. First analysis, the researcher found that all of the test items used communicative language both the main question and the options. Second analysis, the result showed that all of the test items were stated in simple and clear language in both of the main questions and the options. The last analysis, the researcher found that all of the test items were free from non-functional material.

Results from test items validity were divided into content validity, item difficulty, item discrimination, and distractor analysis. For content validity, the researcher found that all of the test items only match with test blueprint. The test items were made without test indicator but only made based on test blueprint. For item difficulty, the researcher found that: 10% of test items were classified as too difficult; 68% test items were classified as moderate; and 22% of test items classified as too easy. The mean of item difficulty the index showed 0.528 means that the test items were in the middle index (moderate) or it was in normal distribution. For item discrimination, the researcher found that: 24 % of test items classified as poor; 44% of test items classified as satisfying; 32% of test items classified as good; and 0% of test items classified as excellent. The mean of item discrimination was 0.317 means that the test items were in the index satisfying. For distractor analysis, the researcher found that: 70% of the test items were classified as test items with good distractor; and 30% of the test items classified as test items with bad distractor.

Third, the test was about the reliability of test items. The result divided into: internal consistency (Kuder Richardson 20), standard error of measurement and source of error. For internal consistency, the researcher used Kuder Richardson 20 (K-R20) method. The result from Master TAP shown that K-R20 index was 0.798 which interpretation was the test items good for a classroom and mark as reliable test reliability. If the teacher wants to obtain K-R20 reliability index 0.8.. , the teacher had to add 51 test items which similar quality with the test right now. To increase it until 0.9.. the teacher should add 114 test items which similar quality with the test right now. For SEm (Standard error of Measurement), the result of analysis divided into two form of result there were: bar graph and item and test analysis. Based on bar graph result, the skewed of score distribution showed positive skewed because most of the score are on the left and the tail form to the right. Based on item and test analysis result, SEm from K-R20 was 3.137. The researcher concludes that SEm index for this analysis is 3 means that the real test score range -3 or +3 SEm. For source of error, the result of analysis divided into: error within test taker, error within the test, error within test administrator, and error in scoring. First, in error within test taker, the researcher found that: there was no student in the Note News;

there was one student who didn't wrote National Examination Participant's Number in their answer sheet; there were two students who wrong in writing National Examination Participant's Number in their answer sheet; there were 23 students who circle more than one answer in their answer sheet; and there were 13 students who answer less than 50 test items in their answer sheet. Second, in error within the test, the researcher divided the results into: potential problem test item and error in test construction. In potential problem test item, the researcher found that there were 21 test items which have marks as potential problem based on master tap result. For error in test construction, the researcher focused on test construction which consists of 21 test items which analyze into 6 categories. The researcher found that: the main questions for all test items formulate clearly; all of the test items did not contain ambiguous meaning both of the main question and the options; there were 4 test items didn't have same length of options; all of the test items were free from incorrect word both of the main question and the options; there were 7 test items which categorized as tricky test items; and there were 9 test items which conclude as test items with too high reading level. Third, in error within the test administrator the researcher only focused on physical comfort for 18 rooms which the students used to do the test. The result can be divided into: all of the room classified as room with good temperature; all of the room classified as room with good lighting because the lighting can be accept from windows and from lamp; there were 5 rooms which have error because of noisy place; and the school follow the rules from Cabinet Minister Regulation to place only 20 chairs and desks for each room. Forth, for error in scoring, the researcher divided the result into: there were 35 students with the percentage 10.17 % who get higher score because of error in scoring and there were 52 students with the percentage 15.11 % who get lower score because of error in scoring.

Discussions

The appropriateness the test items with the principle of test construction in the test for students covered several aspects. For material aspect, the strength was found from all of the test items only had one right answer but for the weakness was found from the test items were written without indicator of competence attainment. According the handbook from Depdiknas (2011), the teachers should make the indicator which referring to Basic Competence and pay attention on context/material chosen. Because this test items were written without indicator of competence attainment, the test items became less specify with the material which being tested. For construction aspect, the researcher found two strengths and four weaknesses in test items. The strengths were: the main questions for all test items formulate clearly and the test items were free from ambiguous meaning. This result suitable with theory from Miller (2008), who stated that multiple choice stems should be free from irrelevant information. The weaknesses were found from: not all of the test items have the same length of options; incorrect words which occur in several test items both the stems and the options; the tricky test items; and reading level too high. For language aspect, the researcher found the strengths in test items. The strengths were all of the test items: used communicative language; stated in simple and clear language and free from non-functional material. This result in line with Depdiknas (2011), stated that test items must use communicative language and not repeat the word or phrase which not part of united concept. All of the test items formulate clearly in form of language and only used material from the test blueprint.

The validity of test items in the test for students covered several aspects. For content validity analysis refers to analysis the suitableness between test blueprints of national examination with test items. Based on the interview with the teachers, the test items were made referring to a handbook from Depdiknas (2011). But in fact, the

teachers only follow the principle of test construction not the guideline of test construction as a package. This condition became the weaknesses that the researcher found in test items. For item difficulty, the mean of item difficulty was 0.528 which means that the result was in moderate index. This result in line with theory from Miller (2008), who stated that most norm-referenced test developers recommend a .30 to .70 difficulty range with an average item difficulty of .50 to maintain a normal distribution. Because the result was at 0.528 it can be conclude that the item difficulty of this test items still in normal distribution and marks as balance for difficulty level. For item discrimination, the result was 0.317 means that item discrimination index was in satisfying index. This result supported by Kubiszyn and Borich (2006), who stated that there was no single answer about good discrimination index, but some experts insist that item discrimination should be at least .30, while others believe that as long as item discrimination in positive value, the ability is adequate. After related the findings and the theory, the researcher conclude that this test items were acceptable as test items with high discrimination index because the index was in positive value and the index still .30. For distractor analysis, the percentage test item with good distractor was 70% (35 items) and the percentage test item with bad distractor was 30% (15 items). According to Kubiszyn and Borich (2006), the bad distractor was the distractor which chosen by more students in the upper group rather than in the lower group. From the theory above, the good distractor was the distractor which chosen by more students in lower group. Because there was about 30% (15 items) which consider as test items with bad distractor, the teachers should modified the distractors to create better items.

The reliability of test items in the test for students covered several aspects. For K-R20 result, the findings showed 0.798 means the test items already good for a classroom and mark as reliable test reliability. Although the reliability already good, the teachers still can increase the reliability by replaced or modified the test items which marks as potential problem. For SEm, the result from K-R20 was 3.137 or 3. The true scores for all students in this test were (score - SEm) – (score + SEm). Although the skewed of score distribution showed positive skewed, the SEm index for this test was large number because the index more than one. It is really suggested for the teachers to carefully review and revised the test. For source of error, the discussions divided into several parts. First, error within the test taker refers to the error cause by the test taker or the students. The researcher found several error which caused by the students. This result in line with the theory from Kubiszyn and Borich (2006), stated that intra-individual error can occur result obtained score lower than student's true score. Second, error within the test, the researcher found that there were 21 test items which mark as potential problem and the weaknesses arose in the potential problem test items. The weaknesses were: length of the options, tricky test items, and reading level too high. Third, error within the test administrator which focused on physical comfort: room temperature, humidity, lighting, noise, and setting arrangement. The findings show good result because the researcher only found one error and it was about the noisy level for the classrooms near the street. The last, error in scoring refers to the students who get higher or lower score rather than their true score. The result show the error occur much in students who got lower score rather than the true score.

CONCLUSIONS AND SUGGESTIONS

Conclusions

For the test construction and the language used, it already fulfill the principle but the weaknesses came from the test items which made without indicator of competence attainment and it was against the principle of test construction. For item difficulty, item discrimination already show good index and can be conclude that, the test items were valid enough in order to use in the classroom but the weakness arise from content validity

which the test items still not really specify with the learning purpose. For K-R20 index showing good index but for SEm index it is obtain large index. The source of error also show the weaknesses which came from error in scoring and cause the accuracy of reliability test score become impair.

Suggestions

Based on the conclusions above, the researcher would share several suggestions to make a better test in the future: to the teachers, it was suggested to make the test blueprint first before writing the test items in order to make the test items more specify and can represent the learning purpose; to the test administrator, it was suggested to prepare the test well in order to avoid miscommunication which can damage students test score by doing the evaluation related the strengths and weakness of previous test; to the school, it was suggested to hold the briefing with the students before the test to make sure students can follow the rule and do the test well which can be held a week before the test and to the school, it also suggested to make well planned test to make sure the teachers have much time to write good test by schedule the activities several months before the test.

BIBLIOGRAPHY

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

Clay, B. (2001). *Is this a trick question? a short guide to writing effective test questions.*Kansas: Kansas Curriculum Center.

Copy of Cabinet Minister Regulation of Education and Culture No 53 Year 2015 About Learning Achievement Assessment by Teacher and Education Unit for Basic and Middle Education. (2015). Indonesia.

Depdiknas.(2011). *Handbook of composing multiple choice test items.*Balitbang.

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

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

Kumar, R. (2011). *Research methodology.* London: Sage Publications Ltd.

Miller, P. W. (2008). *Measurement and teaching.*USA: Patrick W. Miller and Associates.

Ritchie, J., & Lewis, J. (2003). *Qualitative research practice : a guide for social science students and researchers.*London: Sage Publications Ltd.

Young, J.W., So, Y., & Ockey G.J. (2013). *Test guideline for best test development practice to ensure validity and fairness for international English language proficiency assessment.* United Stated of America: Educational Testing Service.