The Quality of Course Subject Test Items In Sundanese Language Education Program

Haris Santosa Nugraha, Temmy Widyastuti and Ade Sutisna
Department of Sundanese Language Education, Universitas Pendidikan Indonesia, Bandung, Indonesia
{harissantosa89, temmy.widyastuti, ade.sutisna}@upi.edu

Keywords: Validity, Reliability, Difficulty, Discriminating, Feasibility.

Abstract: The purpose of this study is to examine the quality of test items of course subjects in the even semester of 2016/2017 academic year in the Department of Sundanese Language Education of Language and Literature Faculty, Universitas Pendidikan Indonesia. The evaluation is based on the validity, reliability, difficulty, and discriminating power as the basis for determining the feasibility of the test items. This research used quantitative approach with descriptive-analysis method. The data source in this study were taken purposively on the subjects of the categories of learning, language, literature, and culture, that used 160 test items of multiple choice questions type. Documentation Study Technique is used in collecting the data. While data processing technique is processed by compiling, tabulation, scoring, and interpretation. The results of this study are as follow: 1) the test items tested level of validity is 59% valid and 41% is not valid; 2) the test items reliability level is mediocre; 3)the test items difficulty level consisted of 58% moderate, 20% easy, 14% difficult, 7% very easy, and 2% is very difficult; 4) the discriminating power distributed to 2% is very good, 10% is good, 43% is adequate, 39% is inadequate, and the remaining 6% is very inadequate; and 5) the test items analysed feasibility level distributed to 67% is feasible, 12% must be revised, and 21% must be changed.

1 INTRODUCTION

Learning evaluation test is one way to measure the achievement of the implemented learning process. By learning evaluation, educators will be able to recognize the advantages and disadvantages of the learning process that has been done. Thus, the results of the evaluation can be used as a benchmark for the improvement of the insufficiency existed. As meant for improving the teaching method, choosing the right learning method, the use of instructional media, and so forth.

The purpose of the learning evaluation test is to obtain information about the lack of the items test, as a "guide" in making improvements (Arikunto, 2013). This can be completed by processing the test results properly. According to Purwanto (in Sridadi, 2002) proper processing of learning outcomes can be done by means of test items analysis and calculate the validity and reliability of the items that aimed to identify the test items used, whether the items are good, mediocre or insufficient.

According to Nurgiantoro (2013) a good and qualified test items is supported by qualified test items, its effectiveness, and its accountability. Quality test items evaluation activities conducted by educators as a means of determining the feasibility of the test items in order to improve the quality of test items that have been made. This activity is conducted as a process of collecting, summarizing, and using the information from students' answers as a basis for making decisions about each assessment (Nitko, 1996). In addition, the purpose of the test items analysis also helps in improving test items evaluation result through revision or disposing the ineffective test items, as well as recognizing the student's diagnostic information about the comprehension of the teaching material already taught. Therefore, to determine the quality estimation of the test items it is necessary to analyse the validity of the test instrument, overall test consistency, difficulty level balance and good discriminating power.

The validity test is usually used in measuring the validity of the test instrument. According to Nurgiantoro (2013) the process of testing the validity of the test instrument is done based on some consideration; (1) validity refers to the feasibility of

Nugraha, H., Widyastuti, T. and Sutisna, A.
The Quality of Course Subject Test Items In Sundanese Language Education Program.
DOI: 10.5220/0007172806580663
In *Proceedings of the Tenth Conference on Applied Linguistics and the Second English Language Teaching and Technology Conference in collaboration with the First International Conference on Language, Literature, Culture, and Education (CONAPLIN and ICOLLITE 2017) - Literacy, Culture, and Technology in Language Pedagogy and Use*, pages 658-663
ISBN: 978-989-758-332-2

interpretation made based on the test result score associated with a particular use rather than the instrument itself. (2) validity is a matter of degree then prevent to think that the test results are valid and invalid. (3), the assessment of the test validity should be related to the intended use of the test results. Based on the above facts, the validation process is done by collecting all evidence that show the scientific basis of the interpretation of the projected score of the test scores results. Interpretation of the test result is used as an attempt to interpret whether the test instrument can be used to measure what should be measured or not (Sugiyono, 2010).

To measure the consistency of the test items with the whole test, it can be done by testing the reliability of the test items. Reliability tests are subject to trust, reliability, consistency, or stability tests based on predefined criteria. A test is believed to be reliable if it always gives the same result when it is tied to the same group at different times or occasions (Anastasi and Urbina, 1997). Based on the above facts it can be concluded that the test reliability is the level or the degree of consistency of the test itself. A test instrument has good reliability when the measuring instrument has a reliable consistency even though it is implemented and tied to the same group at different times or occasions.

A balanced difficulty level is a matter that is not too easy or too difficult. The test items that are too easy caused students unchallenged to try to solve it, causing boredom and blocking students' cognitive development. Conversely, if the test items are too difficult it will cause students desperate or less motivated to try to do the test because it is considered beyond their ability. But even so, it does not mean that those test items cannot be used. This depends on the purpose of using it. Therefore, the proportion of the test items difficulty levels should be considered by analysing the test items index difficulty 0.0 – 1.0 (Arikunto, 2013).

Based on the discrimination power, a good test item is an item that can distinguish between high achiever and low achiever students or among students who have and have not mastered the learning material (Arikunto, 2013). Discriminating power index of each item is usually expressed in proportion. The higher the discriminating power index of the test items, the more capable the test item distinguishes students who have and who have not comprehended the learning material.

Based on the above explanation, testing the quality of test items are very important to do at the Department of Sundanese Education (DPBD) of Language and Literature Faculty (FPBS), Universitas Pendidikan Indonesia (UPI). Based on the results of pre-research observation, the test items used in each semester in DPBD FPBS UPI are periodically updated by lecturers, although the process of updating the questions is done without the process of evaluating the test item primarily. Revitalization of test items is only based on the argument that the questions must be different from the previous test. To a certain extent, it may initiate good quality test items are replaced and insufficient quality test items are kept. This condition may occur because lecturers understanding of quality test of test items is still inadequate.

If this problem is continually occurred, it may cause the evaluation process will not run optimally. The impact is assessment on the students becomes less objective, because the test instrument used may not have good quality. So that, the measurement of the test instrument needs to be performed to examine and to review each item before the items are used. This test is expected to measure the accuracy of the test items, the level of consistency of the test items, the level difficulty of the test items (difficult, medium, or easy), the discriminating level of the test items (very bad, bad, sufficient, and good), and interpretation of the feasibility of the test items (qualified, need to be revised, or rejected to use).

Based on the background above, the problems identified in this study are still lack of empirical and controlled testing of test items in DPBD FPBS UPI. This should be a concern in order to maintain the quality and objectivity of processes and outcomes of lecturing. In order to make the study to be more focused and measurable, the test items analysed in this study are limited. It is only the test items of Final Exam that are grouped into learning, linguistic, literature, and culture subject that carried out in Even Semester of 2016/2017 academic year.

The benefits of this research are to provide information as a reflection for test items makers about several aspects i.e. the accuracy of test instruments, consistency, difficulty, and discriminating power. Those can provide interpretation of the test items quality, whether feasible to be used or not, and as a source of test items planning of learning pedagogy, linguistics, literature, and Sundanese culture at DPBD FPBSD UPI in the future.

2 RESEARCH METHODS

This research uses quantitative approach with descriptive-analysis method. The steps of descriptive method are organized by adopting Stephen's theory (2004) i.e.: (a) collecting factual information in detail and describing the indications in the field; (b) identifying the problems faced; (c) making comparisons; and (d) determining the implications of experience in planning further improvements.

Object of this research is the test items of DPBD FPBS UPI in 2016/2017 academic year. The sample in this study are the Multiple Choices type test items of DPBD FPBS UPI on Even Semester of 2016/2017 academic year. The sampling technique is executed by using purposive sampling technique. The data used as the sample of this study consists of four groups of subjects (MK), namely learning pedagogy, linguistics, literature, and culture. The objects in this study developed naturally and it is not manipulated by researchers and existence of researchers does not affect the dynamics of the object. This research is *ex post facto*, it means that the data is collected after all the events occurred (Azwar, 2004). This is because the student's answers that analysed are the data which has been implemented before.

The instrument used to capture the data used was the analysis program in *Microsoft Excel 2013*. The instrument was developed by the researcher to get a picture that can answer the purpose of this research. Student answers are expected to provide ideas of the test items quality. To obtain good results, before the research instrument used, it was validated by a professional judgment and learning evaluation expert.

Data collection was accomplished by documentation study of the final exam test items on even semester of 2016/2017 academic year by using research instruments that had been prepared. If the research data had been collected, then the data was analysed. Data analysis technique used was descriptive statistical analysis technique. Descriptive statistical analysis is used to analyse all the data by describing the data that has been collected. Then, the collected data was processed by using descriptive statistics that (1) compiled, (2) tabulated, (3) scored and percentage, and (4) interpreted.

The analysis program in *Microsoft Excel 2013* was used to analyse the question items in this study. Excel is a data-solving program commonly called "spreadsheet", because it can be used to process numbers form data or others. There are two ways to process data in *Ms. Excel*, namely (1) through a special statistical auxiliary program and (2) through statistical functions contained in *Excel*. However, because the validity test, reliability, difficulty level, discriminating power, and the feasibility are not available in this program, so this program is made to simplify the data processing by using statistical functions.

3 RESULTS AND DISCUSSION

This section describes the results that have been achieved on study of the test items quality of DPBD FPBS UPI on Even Semester of 2016/2017 academic year. There are four groups of test items tested. Those are learning pedagogy, linguistics, literature, and culture. The test items are tested quantitatively by analysing the validity, reliability, difficulty level, discriminating level, and feasibility of the test items on Odd semester of 2016/2017 academic year.

3.1 The Validity of The Test Items

The validity of the test items is conducted to measure the level of validity of measurement instrument. Valid test items are used to measure what should be measured (Sugiyono, 2010). The purpose of validity test in this research is to recognize the accuracy of measurement instruments in performing its function, so that the test items obtained are the items that are relevant with the purpose of measurement. The support of each test item is expressed in correlation, so that to obtain the validity of a test item the analytical program in *Ms. Excel 2013* is used.

The interpretation of the validity refers to the opinion of Djiwandono (2011) which sorts the level of correlation related to the coefficient of validity. The coefficient of validity index ranges from 0.0 – 1.0. The validity test results from 0.0 – 0.19 are very low, 0.20 – 0.39 are low category, 0.40 – 0.59 are medium category, 0,60 – 0,79 are high category, and 0,80 – 1,00 are very high category.

Based on analysis result the test items of validity test of DPBD FPBS UPI which cover group of learning pedagogy, linguistics, literature, and culture field. The validity levels of test items of DPBD FPBS UPI i.e. 38% are low, 29% are extremely low, 21% are moderate, 11% are invalid, 1% is high, and 0% is extremely high. Base on the above data, the low criteria of validity level has dominant position than others criteria. The criteria are obtained based on *the Product Moment* correlation test on test items

that are interpreted based on the result of comparison between t_{count} value and t_{table}. The greater t_{count} than t_{table} the higher the validity level of a test item, and vice versa. Whereas if the data are grouped based on valid and invalid criteria, then 59% of test items are valid and 41% are invalid. Percentage description of the validity test is based on the correlation of coefficient interpretation of each item. It shows that the test items are still low and cannot measure what should be measured yet. So, based on the explanation above, it shows that the validity of test items of DPBD FPBS UPI on Even Semester of 2016/2017 academic year is generally still in the low category.

3.2 The Reliability of Test Items

Reliability test is performed to identify the consistency level of the answer in the test instrument. This analysis essentially examines the reliability of test items which is a set of questions being given repeatedly on the same object. Good instruments have consistent answers with relatively similar results (Arikunto, 2013). The reliability index ranges from 0.0 – 1.0. The higher the reliability coefficient of a test or it approaches 1.0, the higher the test accuracy. The reliability levels categories are considered very low category is 0.0 – 0.19, low category is 0.20 – 0.39, medium category is 0.40 – 0.59, high category is 0.60 – 0.79, and extremely high category is 0.80 – 1.00.

Based on the analysis results, the test items groups that have high coefficient category is the linguistic group that has 0.75 points and literary groups that has 0.67 points. The test items group with medium category is culture group that has 0.52 points. The problem group with low category is learning pedagogy group that has 0.28 points. The category of reliability found shows that the reliability of the test items is "good", but those are still "need improvement". The reliability includes the accuracy of measurement results, and the stability of the measurement results. So that, if test is conducted several times on this test items, it will give a predetermined result. According to that statement, it is necessary needed a revision of the test items used. So that, the reliability of the test items used is trustworthy.

3.3 The Difficulty Level of Test Items

The level of difficulty (TK) is an index that states the level of difficulty of an item for the test participants. According to Arikunto (2013), the difficulty index is often classified as difficult, moderate, and easy. The difficult question is the question that has difficulty level (P) 0,00 – 0,30, moderate question has (P) 0,31 – 0,70, and easy question has (P) 0,71 – 1, 00.

Based on the result of analysis, the difficulty levels of DPBD FPBS UPI are 57% moderate, 27% easy, and 16% difficult. The proportion of difficulty level distribution is good, because a good test item is a question that ranges from not too difficult to not too easy level. If the test item is too difficult the participant cannot answer the question. If the test item is too easy, so that all students can answer correctly. It means that the test item cannot reflect the students learning achievement. In other words, the test items cannot discriminate between high group test participants (test takers who answered many questions correctly) and low group test participants (test takers who answered many questions incorrectly).

3.4 The Discriminating Power of Test Items

The discriminating power test are performed to determine whether or not the questions can discriminate between high-ability students and low-ability students (Arikunto, 2013). Number that indicates the magnitude of the discriminating power is called the discriminative index (D) which is ranging from 0.00 to 1.00. Questions that categorized as question with extremely bad discriminating power is question which have 0.0 – 0.19 discriminative index, the low category is 0.20 – 0.39, the moderate category is 0.40 – 0.59, the high category is 0.60 – 0.79, and very high category is 0.80 – 1.00.

Based on the results of the discriminating power analysis, there are found five categories that include 43% questions are good enough category, 39% are bad category, 10% are good category, 6% are extremely bad category, and 2% are extremely good. The data shows that the discriminating power of questions tested is dominated by good enough and good category. This indicates that the test instrument is good enough to distinguish high-ability and low-ability students.

3.5 Test items Feasibility

Test items Feasibility is a phase to determine whether the test items analysed are qualified or not, and whether those should be revised or not, or be rejected. The validity test, reliability test, difficulty

level, and discriminating power are used as reference in the process of determining the feasibility of the test items. The test items accepted are those whose criteria of it validity is valid, whose reliability criteria is at least medium, the difficulty criteria is medium/difficult, and the criteria of the discriminating power is good enough/good/extremely good. The test that have to be revised are the test items that have the validity criteria is valid, its reliability criteria is minimal, its difficulty criteria is easy, and its discriminating power criteria is sufficient. While the test items are rejected if the test items validity is not valid, its reliability criteria is low, its difficulty criteria is easy or difficult, and the discriminating power criteria is extremely poor.

Based on the results of test items feasibility analysis of DPBD FPBS UPI at Even Semester in 2016/2017, there are 67% test items are feasible items, 12% must be revised, and 21% must be replaced. Based on the composition of the course subject group, the linguistics group goes to be the group whose questions are the most categorized worthy to be used, followed by culture, literature, and learning pedagogy. The existence of the language learning test items analysis, it has provided a systematic procedure that offers very specific information about the test items prepared. Analysis of this test is conducted as one of the activities that need to be held in order to improve the test instruments quality, both the quality of the overall test and the quality of each test items as part of the test. The test as an evaluation instrument is expected to produce an objective and accurate score. Therefore, it is necessary to make sure that the tests given to the students are as good as possible and good quality.

A good test can be used over and over with a few changes. If there is a test that has poor quality, it will be better the test is discard or not used to test the students. A test can be classified as a feasible measurement instrument, if it fulfils the test requirements. The test requirements that include validity, reliability, have the discriminating power, and have good difficulty level.

4 CONCLUSIONS

Based on the results and discussion above, it can be concluded that the validity level of the test items tested distributed to 59% are valid, while the rest 41% test items are invalid. The number of valid test items is greater than the invalid one. It indicates that many test items tested are generally capable to measure the competence of students that is in line with the course subject content, but rest of them are not capable. The reliability level of the question tested is in the medium category. This shows that the test items have good reliability. The reliability includes the accuracy of measurement results, and the stability of the measurement results. So that, if several tests are conducted on the test items, these will give a predetermined result. Based on the level of complexity, there are about 58% of test items are medium category, 20% test items are easy, 14% test items are difficult, 7% test items are very easy, and 2% test items are very difficult. The result of this test states that the test items tested have a good proportion. Based on the discriminating power, there are 43% test items are considered good enough, 39% are bad, 10% are good, 6% are extremely bad, and 2% are extremely good. The average of the differences level of the test items tested is in good enough to be able to discriminate between the answers of students who have high-ability level and students who have low-ability level. The feasibility level of the test items analysed i.e. 67% are feasible, 12% must be revised, and 21% must be replaced. It is found in this analysis that the majority of test items are qualified to meet the requirements of good quality test items based on validity, reliability, difficulty, and differences power. Based on the above conclusions, the result of the test items analysis at DPBD FPBS UPI reveals good enough result. In spite of this, there are also test items that have invalid status, no discriminating power, disproportionate in the difficulty level, and unacceptable feasibility. Therefore, it needs to revise or to improve the test items of the learning pedagogy groups, linguistics, literature, and culture in further tests. The test items that need to be improved should be selected from the course subject that has been learned and discussed by the student and presented in high quality test items. To realize that, the ability, carefulness, and good experience of lecturers are required to improve the quality of the feasibility of the test items. So that, the results will be more accurate in measuring learning competencies that have been achieved by the students.

REFERENCES

Anastasi, A. and Urbina, S., 1997. *Psychological Testing*, Prentice-Hall, Inc. New Jersey.

Arikunto, S., 2013. *Dasar-dasar Evaluasi Pendidikan*, PT. Bumi Aksara. Jakarta.

Azwar, S., 2004. *Metode Penelitian*, Pustaka Pelajar. Yogyakarta.

Djiwandono, S., 2011. *Tes Bahasa: Pegangan bagi Pengajar Bahasa*, PT Indeks. Jakarta.

Nitko, A. J., 1996. *Educational Assessment of Students*, Merrill an imprint of Prentice Hall Englewood Cliffs. Ohio.

Nurgiantoro, B., 2013. *Penilaian Pembelajaran Bahasa berbasis Kompetensi*, BPFE-Yogyakarta. Yogyakarta.

Sridadi, M. P., 2002. Analisis Butir Soal Pilihan Ganda. *Majalah Olahraga. 8*(26), pp.26-37.

Sugiyono, 2010. *Statistika untuk Penelitian*. Alfabeta. Bandung.