ANALYSIS OF THE QUALITY OF TEACHER-MADE READING COMPREHENSION TEST ITEMS USING ITEMAN

Intan Fitriani Aulia, Muhammad Sukirlan, Sudirman
intanfaulia@yahoo.com
University of Lampung

ABSTRACT

Penelitian ini bertujuan untuk mengidentifikasi: tingkat kesulitan, daya beda, kualitas pilihan, keandalan, keberterimaan; validitas tes dan pendapat guru dan kepala sekolah tentang penggunaan *iteman*. Objek penelitian adalah soal mid semester bahasa Inggris yang guru buat untuk kelas 2 SMPN 8 Bandar Lampung tahun 2013/2014. Hasil penelitian menunjukan: tingkat kesulitan terdiri dari 21 item (52.5%) diterima, 5 item (12.5%) perlu diperbaiki, 2 item (5%) perlu dibuang. Daya beda terdiri dari 16 item (40%) diterima, 2 item (5%) perlu perbaikan, dan 17 item (42.5%) perlu dibuang. Kualitas pilihan (Prop. Endorsing) adalah 42 pilihan (26%) diterima, 35 pilihan (22%) perlu diperbaiki dan perlu dibuang. Keandalan (alpha) adalah 0.763. hasil juga menunjukan bahwa guru tidak pernah menganilisis item tes sebelum mereka menggunakannya karena mereka belum begitu mengenal perangkat *iteman* dan karena keterbatasan waktu.

The objectives of the research are to identify: the difficulty level, discriminating power, quality of the options, reliability, acceptability; validity of the items, and the teacher's as well as headmaster's opinions about the *iteman*. The object of the research is teacher-made English Mid Semester for second grade in SMP Negeri 8 Bandar Lampung in 2013/2014 academic year. The results show that: difficulty level consists of 21 items (52.5%) acceptable, 5 items (12.5%) need revising, and 2 items (5%) need dropping. Discriminating power consists of 16 items (40%) acceptable, 2 items (5%) need revising, and 17 items (42.5%) need dropping. The quality of the options is 42 options (26%) acceptable, 35 options (22%) need revising and need dropping. Reliability is 0.763. It also shows that the teachers never analyze the test items before using them because of their unfamiliarity with *iteman* and the time constaint.

Keywords: discriminating power, *iteman* software, difficulty level, reliability, and validity.

INTRODUCTION

The importance of assessment is to make sure that the objectives of the teaching and learning have been achieved. To achieve the purpose above, the teachers have to make sure that all the test items that they use to assess their teaching and learning processes in the classroom are of a good quality. To analyze teacher-made exam, the teachers usually use the manual method. Actually, the teachers can analyze the item of the test using software. One of the best software that can be used to analyze the test items is *iteman* software. The problem concerning with the analysis of test items is very important to be investigated because all teachers who teach English in SMPN 8 should be able to tryout all the test items that they make before they use them to test the students' mastery of the material that they have learned.

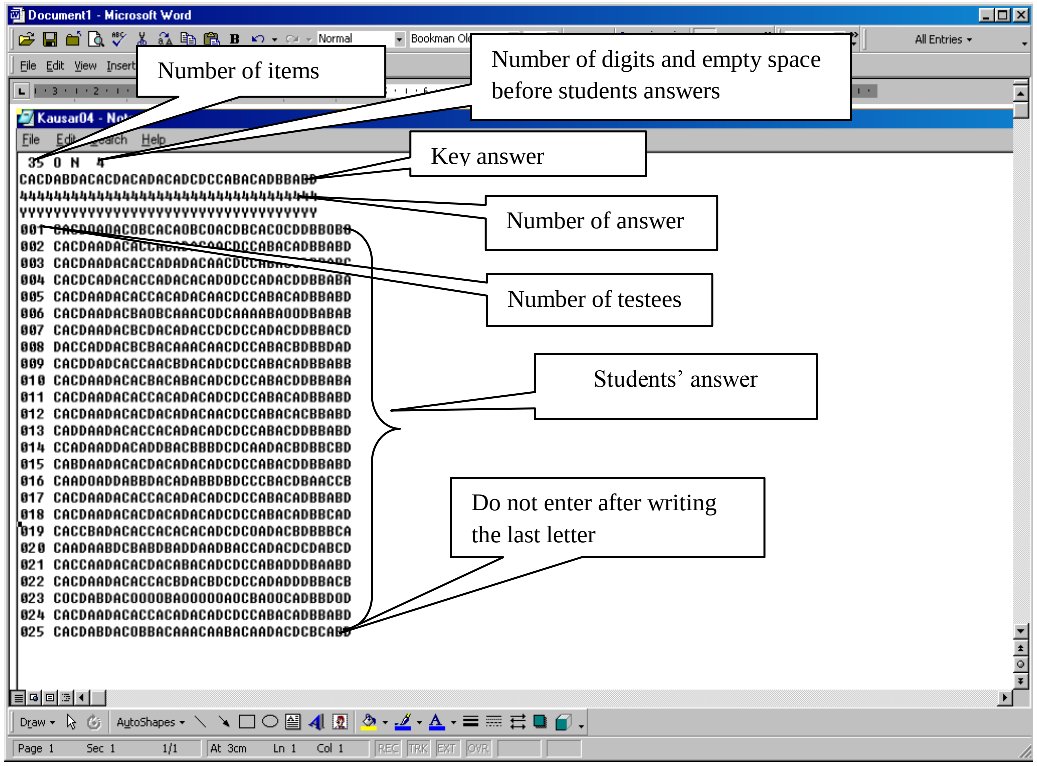
The objectives of this research are to analyze and to solve the following issues: The level of difficulty of the test items in reading that the teachers made for Mid Semester, the discriminating power of the test items the teachers made for Mid Semester, the qualities of the options of the test items that the teachers made for Mid Semester, the reliability of the test items that the teachers made for Mid Semester, the validity of the test items that the teachers made for Mid Semester, the number out of 40 items are acceptable/revised/dropped, and to investigate the teachers' opinion about *iteman*.

ITEMAN SOFTWARE

Iteman is software to analyze test item which is intended to determine which test item is good and which is not, based on the criteria of a good test item which consist of validity, reliability, discriminating power, and level of difficulty. According to assessment system corporation ASC (1989-2006) as Suparman (2011: 86) quoted, *iteman* can be defined as one of the analysis programs consisting of assessment system corporation's item and test analysis package. *Iteman* is very important for teachers of English in all levels (Junior High and Senior High School) who have to be responsible for administering tests, such as mid semester or final examination, so that they can be sure about the quality of the test item that they will use.

In order that the teachers can make use of *iteman* effectively, the teachers have to know how *iteman* works. According to Suparman (2011: 86), the data, which have been input or keyed into the computer to be analyzable by *iteman,* should be formatted in ASCII (text only) files. The keying of the input data, according to him, can be completed successfully using three components: a) the *iteman* for windows text editor, b) note pad, and c) a word processing editor that produces ASCII output. He further states that it is very important to note that all the data to be included in the analysis must be contained in a single input file. One of the advantages of the *iteman*, according to Suparman (2011: 86), is that a single analysis can accommodate up to 750 items, while the number of the examinees is almost unlimited.

There are five primary components to put a data file in an *iteman*: 1) a control line describing the data, 2) a line of keyed responses, 3) a line of the numbers of alternatives for the items, 4) a line specifying which items are to be included in the analysis, and 5) the examinee data, (ASC, 1989-2006: 2) Suparman (2011: 86) quoted. The following is an example of a data file on an iteman:


Document1 - Microsoft Word
File Edit View Insert
Kausar04 - Not
File Edit Search Help
Number of items
Number of digits and empty space before students answers
Key answer
Number of answer
Number of testees
Students' answer
Do not enter after writing the last letter
Draw AutoShapes
Page 1 Sec 1 1/1 At 3cm Ln 1 Col 1 REC TRK EXT OVR

```
 35 O N  4
CACDABDACACDACADACADCDCCABACADBBABD
44444444444444444444444444444444444
YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY
001 CACDOADACOBCACAOBCOACDBCACOCDDBBOBO
002 CACDAADACACCACADACAACDCCABACADBBABD
003 CACDAADACACCADADACAACDCCABACADBBABC
004 CACDCADACACCADACACADODCCADACDDBBABA
005 CACDAADACACCACADACAACDCCABACADBBABD
006 CACDAADACBAOBCAAACODCAAAABAOODBABAB
007 CACDAADACBCDACADACCDCDCCADACDDBBACD
008 DACCADDACBCBACAAACAACDCCABACBDBBDAD
009 CACDDADCACCAACBDACADCDCCABACADBBABB
010 CACDAADACACBACABACADCDCCABACDDBBABA
011 CACDAADACACCACADACADCDCCABACADBBABD
012 CACDAADACACDACADACAACDCCABACACBBABD
013 CADDAADACACCACADACADCDCCABACDDBBABD
014 CCADAADDACADDBACBBBDCDCAADACBDBBCBD
015 CABDAADACACDACADACADCDCCABACDDBBABD
016 CAADOADDABBDACADABBDBDCCCBACDBAACCB
017 CACDAADACACCACADACADCDCCABACADBBABD
018 CACDAADACACDACADACADCDCCABACADBBCAD
019 CACCBADACACCACACACADCDCOADACBDBBBCA
020 CAADAABDCBABDBADDAADBACCADACDCDABCD
021 CACCAADACACDACABACADCDCCABADDDBAABD
022 CACDAADACACCACBDACBDCDCCADADDDBBACB
023 COCDABDACOOOOBAOOOOOAOCBAOOCADBBDOD
024 CACDAADACACCACADACADCDCCABACADBBABD
025 CACDABDACOBBACAAACAABACAADACDCBCABD
```

Fig1. An example of Data file using notepad on Windows

Source: Suparman (2011: 87)

Steps of Entering the Data Using a New File

The *iteman* program can function only with multiple-choice item. To analyze test item using the *iteman* program is relatively easy. The most important thing to do is to be careful in keying the data into the computer because if the data are inputed wrongly it will result in imprecise finding of the data analysis. Suparman (2011: 87) considers that an item analysis should use nine steps to enter the data using a new file as follows:

1. Click **Start**
2. Select **program**
3. Select **accessories**
4. Choose and click **Notepad**
5. Save/ click **file**

6. Select and click **save as**, then name the data file, for example: Advread (make sure the file name must not exceed eight letters/ numbers.
7. Start data entry, it will be faster if you work with your friend- one of you reads students' answers and the other types them. If you work with your friend, please make sure to pronounce the letter clearly, e.g., *a* for *apple*; *b* for *ball*; *c* for *Charlie*; *d* for *doctor*; and *e* for *ent*.
8. It is advisable for you to save it frequently by clicking **File** and the **Save** so that the typed data will not loss if the electric current suddenly cuts off.
9. The data will appear like shown on the Fig 1above.

In the following paragraph, the steps of how to analyze the data using *iteman* program is put forward. There are six steps that have to be done by the item analysis as follows:

1. Open *iteman* program, by clicking **Star**t,
2. Select program/ click *iteman*
3. Type the name of your data file (input) as you like on *Enter the name of the input file*. For example **F:\advread.txt then Enter**
4. Enter the name of the output file on *Enter the name of the output file*. For example, in this case: F:\advread.output then click **Enter**
5. A question will appear. **Do you want the scores written to a file?** (Y / N), then type **Y** and click **Enter**.
6. Enter the name of your score file on *Enter the name of the score*: for example, F:\Advread.scr. Then click **Enter**. Finish.

The Results of Data Analysis

The results of the data analysis in this research were based on the formulation of the problems as stated and based on the output data of the *iteman*. That is, the analysis of the students' answers to the test items used in mid semester examination in SMP Negeri 8 Bandar Lampung. The data in the output of *iteman* is shown in the following figure, Figure 4.1 (see the next page).

As the Figure 4.1 shows, there are two statistics that is provided by the *iteman*, that is, *Item Statistics* and *Alternative Statistics*. Each of them has each own components comprising: seq. *no, scale-item, prop. correct, disc. index,* and *point biser*. Whereas *Alternative Statistics* consists of the following components, that is, *prop. total, endorsing low, endorsing high, point biser,* and *key*. These components are shown by the following figure.

```
                 Item Statistics                  Alternative Statistics
            -----------------------          ---------------------------------
Seq.  Scale   Prop.   Disc.   Point            Prop.  Endorsing   Point
No.   -Item  Correct  Index   Biser.    Alt.   Total  Low   High  Biser. Key
----  -----  -------  ------  ------    -----  -----  ----  ----  ------ ---

  1    0-1     .24      .50    .65        A     .57   .60   .50   -.25
                                          B     .05   .20   .00   -.23
                                          C     .24   .00   .50    .65    *
                                          D     .14   .20   .00   -.30
                                        Other   .00   .00   .00


  2    0-2     .52      .63    .57        A     .48   .80   .17   -.57
                                          B     .52   .20   .83    .57    *
                                          C     .00   .00   .00
                                          D     .00   .00   .00
                                        Other   .00   .00   .00

  3    0-3     .52      .27    .38        A     .48   .60   .33   -.38
                                          B     .00   .00   .00
                                          C     .52   .40   .67    .38    *
                                          D     .00   .00   .00
                                        Other   .00   .00   .00
```

```
 4    0-4      .38       .63    .57        A      .52   .40   .17   -.39
                                           B      .05   .20   .00   -.19
                                           C      .38   .20   .83    .57   *
                                           D      .05   .20   .00   -.19
                                         Other    .00   .00   .00

 5    0-5      .67       .23    .09        A      .67   .60   .83    .09   *
                                           B      .14   .00   .17    .13   ?
          CHECK THE KEY                    C      .05   .20   .00   -.19
  A was specified, B works better          D      .14   .20   .00   -.13
                                         Other    .00   .00   .00
```

And so on.

Figure 4.1 Item Statistics and Alternative Statistics

Quality of Item Test

a. Concept of Validity

One of the most important characteristics of the test is validity. Validity is very important because a testing will gain nothing if the test is not valid for the use that we want to make of it. However, a test which may have a high validity for one purpose may have only moderate validity for another, even it may be to slight to be of important. Validity is divided into four types: face validity, content validity, construct validity, and criterion related validity.

1) Face Validity

Face validity refers to the states that the test looks as if it should be valid. According to Lyman (1971: 21-23) good face validity may help to maintain motivation high because people tend to try if the test look reasonable. However, face validity is not as important as other indications of validity.

2) Content Validity

Content validity is more systematic and more sophisticated. It is none as logical validity, cost validity, curricular validity, or textbook validity. Similar to face validity, content validity is not statistical. Content validity is very important in achievement test. According to Hatch and Farhady (1982: 250-251) content validity can be defined as the extent to which a test measures representative sample of the subject meter content. The adequacy of the sample and not only on the appearance of a test is the focus of content validity. The content of whatever material we want to measure must be carefully defined.

3) Construct Validity

The third category is construct validity. It is the most important type in psychological theory. Principally, contract validity deals with the psychological meaningfulness of the test. With construct validity, the results can be predicted which logically should be obtained if the test is valid. According to Lyman (1971:23) the prediction is stated concretely enough and precisely enough, consequently it can be tested statistically.

There are many psychological constructs that are very important dealing with success in language learning. A construct is like *self-esteem extrovert*, *acculturated*, and *motivated* are the task of establishing an index construct validity.

4) Criterion Related Validity

Criterion related validity is the validity of the test when it is used to predict future performance or to estimate current performance based on some valued measures other than the test itself. For example, a language aptitude test has been pretended and it is thought that it is good one. After that, the test is administered to a group of beginning language learners, and to show it is a valid test, the results of the test should be compared with an established test, such as TOEFL, which is the criterion that is expected to be able to predict. It is predicted from the aptitude test course to performance on TOEFL.

However, if two tests such as TOEFL and ALIGU are administered at the same time and compared, it is checking concurrent validity. Criterion related validity is the criterion for the other. Hatch and Farhady (1982: 251) define criterion related validity as the extent to which the test performance is related to some other valued measure of performance. The validity is the degree to which the first test, for example a teacher made test, is seen as related to established criterion.

b. Concept of Reliability

Test reliability is very important for a test user because it is necessary for good validity. In short, a test can be highly reliable without necessary being valid for any purpose of interest. *Test reliability* refers to the reproduced ability of test results. In short, a test with high reliability is one that will reproduce very much the same relative important of test score for a group of students under different conditions or situations.

c. Discriminating Power

Discriminating power refers to the capacity of a test to discriminate between the clever and the stupid students. There are two indicators of the item discrimination effectiveness, which a point biserial correlation and biserial correlation coefficient (Matlock-Hetzel, 1997). The advantage of using discrimination coefficient over the discrimination index (D) is that every person taking the test is used to compute the discrimination coefficient and only 54% (27% upper + 27% lower are used to compute the discrimination index, (D). The point biserial (rpbis) correlation is applied to determine whether the right people are getting the items right, and how much predictive power the item has and how it would contribute to prediction.

To measure discriminating power (D), three ways can be used: a. discriminating index; b. correlation index; and c. harmonious index (Suparman, 2011). A discriminating power is usually symbolized with a capital D. There are several steps that should be followed to determine the level of discriminating power: first, rank order the answer sheets top-down from the highest to the lowest scores based on the total number of test takers; then multiply N with 27%, the result is N score; after that, calculate N for the upper group (the answer sheets with high scores are counted from the top). While N for the lower group (the answer sheets with low scores are counted from the bottom). And finally, determine the proportion of the test items answered correctly by each group. That is, the correct answers from each of the Upper Group (UG) and lower Group (LG) are divided by N. The discriminating power is in fact the differences of the proportion of the correct answers between the UG and the LG. So, it can be stated that D = UG – LG.

There are three parametric criteria that the teachers can follow to determine whether a test item is accepted, revised, or rejected, as follows:

Parameter of D Coefficient	**Decision**
D = > 0.30	Accepted
D = 0.10 – 0.29	Revised
D = <0.10	Rejected

Source**:** Suparman (2011: 93)

d. Level of Difficulty

Level of difficulty is simply the percentage of students taking the test who answered the item correctly. The larger the percentage getting an item right, the easier the item. The higher the difficulty index, the easier the item is understood to be (Wood, 1960). According to Crocker and Algina (1986: 93) the difficulty index of a test item tells a teacher about the comprehension of or performance on material or task contained in an item.

The difficulty level of an item is known as index of difficulty. Index of difficulty is the percentage of students answering correctly each item in the test. Index of discrimination refers to the percentage of high-scoring individuals responding correctly versus the number of low-scoring individuals responding correctly to an item. This numeric index indicates how effectively an item differentiates between the students' who did well and those who did poorly on the test.

METHOD

The design of current research is descriptive and evaluative, that is, the research describes the result of an evaluation on an object which is based on standard criteria. The object in this research consists of the test items and the students' answers to the test. Both of test items and the students' answers are analyzed using standard criteria, that is, level of difficulty, discriminating power, qualities of the options, reliability, and validity.

The current research is carried out at SMP Negeri 8 Bandar Lampung, second grade on semester three, in 2013/2014 academic year. This research is carried out for three months. The activities are consist of preparing the proposal, logging it, determining the object of the research, determining the subjects, approaching the school and teachers, seeking for permission from the head master and teachers of English to carry out the research in that school.

To collect the data there is two data collecting techniques that are used, that is, test that the teachers made for mid semester and interview. The test is used to gather students' answers to the question, whereas the interview is organized to trace the teachers of English experience and opinions about *iteman* and its implementation.

DATA ANALYSIS

The analysis is using the following procedure of how to interpret the results of the item analysis. The procedure has been summarized based on the recommendation of some experts of measurement:

Table 3.1 Criteria of Test Item Quality

Prop Correct (Level of Difficulty – p)	
0.000 - 0.250	Difficult
0.251 – 0,750	Average
0.751 – 1.000	Easy
Point Biseral (Discriminating Power – D)	
0.199 -	Very low ≤ D
0.200 – 0.299	Low
0.300 – 0.399	Average
0.400	High ≥ D
Prop Endorsing (Proportion of the Answer)	
0.000 – 0. 010	Low
0.011 – 0.050	Sufficient
0.051 – 1.000	Good
Alpha (Test Item Reliability)	
0.000 – 0.400	Low
0.401 – 0.700	Average
0.071 – 1.000	High

Furthermore, the criteria above is necessary for the assessor or teacher to have a guideline to classify each item whether it should be revise, dropped, or can be use directly without any revision. For that purpose, the following guideline can be considered as one of the reference:

Table 3.2 Criteria to classify the quality of test items

Level of Difficulty (p)	
0.000 - 0.099	Very difficult/needs total revising
0.100 – 0,299	Difficult/needs revising

0.300 – 0.700	Average/good
0.701 – 0.900	Easy/needs revising
0.901 – 1.000	Very easy/needs dropping/total revising
Point Biseral (Discriminating Power – D)	
0.199 -	Very low ≤ D/needs dropping or total revising
0.200 – 0.299	Low/needs revising
0.300 – 0.399	Quite average/without revision
0.400	High ≥ D/very good
Prop Endorsing (Proportion of the Answer)	
0.000 – 0. 010	Least/drop, or needs revising
0.011 – 0.050	Sufficient/good enough
0.051 – 1.000	Very Good
Alpha (Test Item Reliability)	
0.000 – 0.400	Low/not sufficient
0.401 – 0.700	Average/sufficient
0.071 – 1.000	High/Good

RESULTS AND DISSCUSSION

Based on the results of the data analysis and discussion, the following conclusions are drawn:

1) The level of difficulty of teacher-made Mid Semester test items can be classified into four categories: *good items*, *very difficult*, *very easy*, and *too difficult*.
2) The discriminating power of the teacher-made Mid Semester test items are classified into four categories, as follows: *high discriminating power*, *quite average*/*without revising*, *low/need revising*, and *very low/need.*
3) The qualities of the options (*Prop. Endorsing* in *iteman* terms) in teacher-made Mid Semester test items in SMPN 8 Bandar Lampung, based on the *iteman* analysis, are classified into three classifications: *need revising*, *good enough*, and *very good*.
4) The reliability of the teacher-made Mid Semester test items based on scale-statistics as a part of *iteman* analysis is categorized as high/good.

5) Concerning the validity of the teacher-made Mid Semester test items, it can be considered moderate validity seen from content validity. This is identified not by means of *iteman* analysis because the *iteman* analysis does not cover the validity. But the validity of the test items should be consulted and compared with the curriculum or syllabus of the school.
6) In general, level of difficulty consists of 21 items (52.5%) is acceptable, 5 items (12.5%) need revising, and 2 items need dropping. Discriminating power (Point Biser) consists of 16 items (40%) is acceptable, 2 items (5%) need revising, and 17 items (42.5%) need dropping. Quality of options (Prop. Endorsing) is 42 options (26%) is acceptable, 35 options (22%), and 35 options (22%). Reliability (alpha) is 0.763.
7) Based on the interview with the teachers and the headmaster it can be concluded that the teachers never analyze the test items before using them because they have not been familiar with *iteman* and because of limitation of the time, that is, the interval between the test items preparation with the administration of the test is very short (one month).

CONCLUSION

Based on the results of the data analysis and discussion, the conclusions are drawn:

level of difficulty consists of 21 items (52.5%) is acceptable, 5 items (12.5%) need revising, and 2 items need dropping. Discriminating power (Point Biser) consists of 16 items (40%) is acceptable, 2 items (5%) need revising, and 17 items (42.5%) need dropping. Quality of options (Prop. Endorsing) is 42 options (26%) is acceptable, 35 options (22%) need revising, and 35 options (22%) need dropping. Reliability (alpha) is 0.763.

Based on the interview with the teachers and the headmaster it can be concluded that the teachers never analyse the test items before using them because they have not been familiar with *iteman* and because of limitation of the time.

SUGGESTIONS

In line with the conclusions above, the following suggestions are recommended:

1. The teacher of English should be trained to be familiar with and use the *iteman* so that they can improve the quality of the test they use, which in turn, can improve the quality of their teaching.
2. The teachers should be familiar with all the terms related to the quality of test items, such as validity, reliability, prop. Correct (level of difficulty), point biserial (discriminating power), prop. Endorsing (options), distracters, key answers, alpha, and standard deviation.

REFERENCES

Crocker, L., and Algina, J. 1986. Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

Hatch, E and Farhady, H. 1982. *Research design and statistics for applied linguistics.* Rowley, Massachusetts: Newbury House Publishers, Inc.

Lyman, H.B. 1971. *Test scores and what they mean. Second ed.* New jersey: Prentice Hall, Inc.

Matlock-Hetzel, S. 1997. *Basic Concepts in Item and Test Analysis.* Texas: A&M University

Suparman, U. 2011. The implementation of *iteman* to improve the quality of English test items as a foreign language: An assessment analysis. *AKSARA - Jurnal Bahasa, Seni, dan Pengajarannya,* Vol-XII, No. 1, pp 86-96.