English Focus
Vol. 3, No. 1,2019
p-ISSN: 2614-638X ; e-ISSN:2614-6355
Website: http://englishfocus.upstegal.ac.id/efj/
DOI:10.24905/efj.v3i1.69

Items Analysis of the Achievement Tests in EFL Classrooms

Adinda Putri Nurbaeti[1], Taufiqulloh[2], Anin Eka Sulistyawati[3*)]

[1,2,3]Universitas Pancasakti Tegal

*) Corresponding author: Email: aninekas@gmail.com

ABSTRACT

This study focused on three kinds of item analysis towards the test items of the achievement tests. The validity and reliability were also provided as supporting functions. This study used quantitative data for the data source and some qualitative explanation to elaborate on the data. To gain the data analysis, the test papers and students answers sheets were collected from three achievement tests of SHS X, SHS Y, and SHS Z. Also, the first-grade students of those schools were as the sample of this study. The study revealed (1) the mean of item facility of three achievement tests categorized as medium test items (SHS X= 0,69; SHS Y= 0,55; and SHS Z= 0,44), while the mean of item discrimination of SHS X examined as good items (0,326) and the mean of item discrimination of SHS Y, SHS Z analyzed as satisfactory items (SHS Y= 0,245; and SHS Z= 0,244). Moreover, half of the distractor efficiency of those tests were accepted. Also, the validity and reliability of the achievement tests were found. Thus, it can be summarized that the achievement tests need to be improved since there are some items have high item facility and low item discrimination.

Keywords:*items analysis, achievement test, senior high school*

INTRODUCTION

The teachers have some roles in the teaching-learning process as the learner, the administrator, the facilitator, the manager, and as the evaluator (Kumbakonam & S, 2017: 1). In line with that statement, the teachers should not only focus on what method the teachers use or which methods suitable for the particular material, but it is also about how the teachers evaluate the students' ability during or at the end of the term of study. Thus, evaluation aims to make such a decision regarding students' ability, knowledge, or performances, through the systematic evaluation in educational programs. And in order to evaluate the students, the teacher will need a test as a tool of evaluation. Since evaluating the students' achievement is not an easy thing, teachers should pay attention on what evaluation method is appropriate for the instrument. Therefore, the teachers administer a test in order to evaluate the students' understanding and achievement towards the material.

Also, further analysis in this study is to find out the validity and reliability index of

each achievement test. Capkova et al. (2015: 2) define that test validity is a measure of how accurately a test score reflects students' real-life language ability. Each test item can be identified as a valid item if the item does measure what the test is supposed to measure. In addition, Tambunan in Ciptaningrum (2014: 8) explains that validity deals to the extent to which the result of an evaluation procedure serve the particular uses for which they are intended. Thus, the validity of the test is the level of validity in which the test measures what is expected to measure.

From the explanation above, it can be concluded that validity is one of the criteria to identify whether the test shows a good test. Therefore, the sentence of measure what is intended to measure means that a good test should measure students' ability and knowledge based on their understanding level. Thus, the relevance of the material and the blueprint of the test is required. Besides that, Grant et al. (2006: 7) explain that reliability is concerned with the consistency of the results produced by the assessment instrument. It is a measure of the extent to which the test scores are free from errors of measurement. Theoretically, a reliable test should produce the same result if administered to the same student on two separate occasions, provided the conditions are the same and there is an adjustment for prior learning and growth. A set of the test could be qualified as a reliable test if they are dependable and consistent. Facilities, human error, environment, and/or students' condition can be factors of the measurement errors.

While, authenticity is the fourth criterion. It could be employed in the following ways, specifically are the nature of language in the test, contextual items, interesting and meaningful topics, some thematic organization to items are provided, such as using a storyline or episode, and tasks represent, or closely approximate, real-world tasks. The explanation above shows that the ways how the test delivered are important. It means that the given language and the features of the target language task should be relevant to the students. Thus, the students will do the test maximally.Last but not least, Washback. It enhances a number of basic principles of language acquisition: intrinsic motivation, autonomy, self-confidence, language, ego, interlanguage, and strategic investment, among others. It can be concluded that the important thing after the test administered that is students deserve to get feedback from the teacher. It may be some praise, constructive criticism, notes or comments. Thus, the students will discover their strength or weaknesses in the material of the test.

This research aims at analyze on three kinds of item analysis towards the test items of the achievement tests.

REVIEW OF RELATED LITERATURE

Here, Brown (2004: 3) states the definition of the test. He states that a test is a method of measuring a person's ability, knowledge, or performance in a given domain. Moreover, Braun et al. (2006: 13) denotes that the term "test" refers to an instrument of assessment that is conducted under some set of formal conditions. This means that the test is administered depending on the school's current regulation.Hence, there are some test types available depend on the purpose of the test itself. H. D. Brown (2004: 43) defines that there are five kinds of tests based on the specific objectives and purposes, namely language aptitude test, proficiency test, placement test, diagnostic test, and achievement test. For this reason, the appropriate test for evaluating the students' performance is the achievement test. However, Capkova, Kroupova, & Young (2015: 2) state that achievement tests are designed to show that students have learned what they have been taught. Similarly, Brown (2004: 47) explains that achievement tests are or should be limited to particular material addressed in a curriculum within a particular time frame and are offered after a course has focused on the objectives in question. Thus, the result of the test can be covered which contains the students' understanding in forms of score.

Related to the importance of item analysis, there are three kinds of the analysis: (1) item facility, (2) item discrimination, and (3) distractor efficiency. **Item facility** is used to know whether the test items are easy or difficult for the students. In line with the sentence, then Wood in Marie &Edannur (2015: 3) stated that the item facility of an item is understood as the proportion of the persons who answer a test item correctly. To calculate IF, add the number of students who correctly answered a particular item and divide that sum by the total number of students who took the test. The second is **item discrimination**, based on H. D. Brown (2004: 59) that item discrimination is the extent to which an item differentiates between high- and low-ability test-takers. In line with that statement, Cohen et al. (2007: 422) discussed that item discriminability, or item discrimination refers to the potential of the item in question to be answered correctly by those students who have a lot of the particular quality that the item is designed to measure and to be answered incorrectly by those students who have less of the particular quality that the same item is designed to measure. The last of items analysis is **distractor efficiency**, based on Mozaffer& Farhan Jaleel in Ciptaningrum (2014: 22) that another important technique is an analysis of distractors, that presents information depending on the individual distractors and the key of each test item. Moreover, According to Fulcher (2010: 173) that distractor analysis

involves counting how many test-takers selected each distractor to discover which are not working as intended. It is can be inferred that the distractor efficiency is important because it shows the students' abilities and performances, especially depending on the high-ability students and low-ability students through their choices so that the teachers may analyze whether the test items refer as good or bad test items. The test items are good if most of the high-ability students are answer correctly, and the little of low-ability students are answer correctly, vice versa.

Tests as a tool of evaluation should be tested to measure whether those tests have fulfilled the criteria as a good test or not. Consequently, H. D. Brown (2004: 19) defined that there are five criteria for testing a test, they are practicality, reliability, validity, authenticity, and washback. A test should be practical to represent as an effective test which can be referred by identifying that the number of the test is consistent with the time estimations, examining a set of test to be easy to organize in the classroom, and making definitely score procedure.

Also, further analysis in this study is to find outthe validity and reliability index of each achievement test. Capkova et al. (2015: 2) define that test validity is a measure of how accurately a test score reflects students' real-life language ability. Each test item can be identified as a valid item if the item does measure what the test is supposed to measure.In addition, Tambunanin Ciptaningrum (2014: 8) explains that validity deals to the extent to which the result of an evaluation procedure serve the particular uses for which they are intended. Thus, the validity of the test is the levelof validity in which the test measures what is expected to measure. From the explanation above, it can be concluded that validity is one of the criteria to identify whether the test shows a good test. Therefore, the sentence of measure what is intended to measure means that a good test should measure students' ability and knowledge based on their understanding level. Thus, the relevance of the material and the blueprint of the test is required. Besides that, Grant et al. (2006: 7) explain that reliability is concerned with the consistency of the results produced by the assessment instrument. It is a measure of the extent to which the test scores are free from errors of measurement. Theoretically, a reliable test should produce the same result if administered to the same student on two separate occasions, provided the conditions are the same and there is an adjustment for prior learning and growth. A set of the test could be qualified as a reliable test if they are dependable and consistent. Facilities, human error, environment, and/or students' condition can be factors of the measurement errors.

While, **authenticity** is the fourth criterion. It could be employed in the following

ways, specifically are the nature of language in the test, contextual items, interesting and meaningful topics, some thematic organization to items are provided, such as using a storyline or episode, and tasks represent, or closely approximate, real-world tasks. The explanation above shows that the ways how the test delivered are important. It means that the given language and the features of the target language task should be relevant to the students. Thus, the students will do the test maximally.

Last but not least, **Washback**. It enhances a number of basic principles of language acquisition: intrinsic motivation, autonomy, self-confidence, language, ego, interlanguage, and strategic investment, among others. It can be concluded that the important thing after the test administered that is students deserve to get feedback from the teacher. It may be some praise, constructivecriticism, notes or comments. Thus, the students will discover their strength or weaknesses in the material of the test.

METHOD

The quantitative data is used as the data source of this study and some qualitative explanation are provided as the supporting data. The purposive sampling is used to gather the data source, Cohen et al. (2007: 114). Therefore, this study uses the result of achievement tests from the first-grade students of Senior High Schools in Pemalang in academic year 2018/2019 as the subject of this research. The data are taken from the three different schools purposively as the sample of the research because those schools administer the 2013 Curriculum, and definitely, they provide the subject *Bahasa InggrisPeminatan*. Thus, the subject in this research is English achievement tests of the Tenth Grade which have been conducted in the first semester of Academic Year 2018/2019 at the three Senior High Schools in Pemalang, especially the test papers or score recapitulation, and students' answer sheets from those three schools which the whole number of test items are 135 from 104 students of those schools.

The first steps of analyzing the data are getting the data from the specific schools in forms of the test-papers and students' answer sheets. Then, the students of each school were split into three groups depending on the score. Those groups are called as the higher achiever students, medium achiever students, and low achiever students. Thus, the item discrimination can be analyzed by using the total of correct answers of high achiever students and low achiever students as the formula. Conversely, the item facility and distractor efficiency do not use those groups as the formula. The item facility does only use

the result of students' correct answers of each item, while the distractor efficiency does not use the mathematical operations in analyzing the data. In addition, validity and reliability are also found in this study. Ultimately, the result of the items analysis, validity and reliability are explained based on the data analyzing.

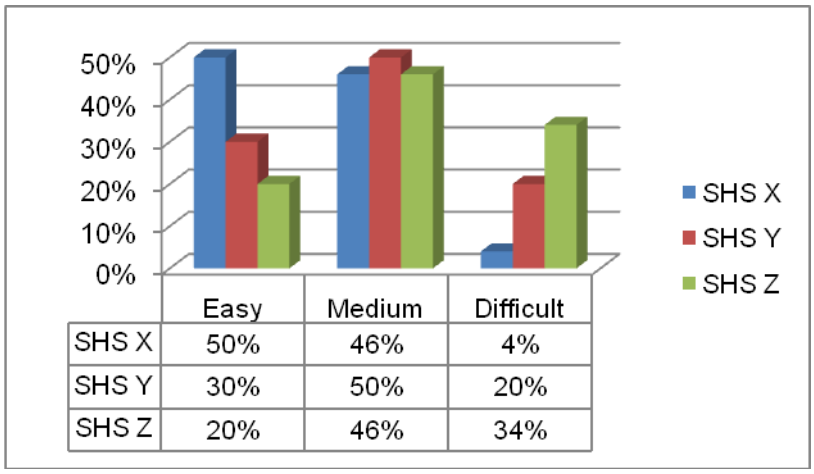
FINDINGS AND DISCUSSION

The procedure of data collection began with some fields analysis have been done by the researcher at the Senior High Schools in Pemalang, particularly the first grade of those Senior High Schools from March till June 2019. Then the researcher gained the data of 135 items of multiple choices from 104 students of three Senior High Schools. However, those data found contained the items analysis for instances the item facility, item discrimination, distractor efficiency. Also, the validity and the reliability supporting functions in this research. In getting the data, the researcher asked for the curriculum unit of those schools to get the result of the test in detail.

The findings of this study focused on the results of items analysis towards the achievement tests. However, making a comparison among each achievement tests based on the items analysis is the procedure in delivering the findings.

1. Item Facility

The item facility of SHS X shows 50% of total test items are easy, 46% of items are identified as medium test items, and 4% of the test items are difficult. While the achievement test of SHS Y proved that 30% of test items are easy, 50% of items are medium, and 20% of items are classified into difficult items. Conversely, the result of item facility towards the achievement test of SHS Z is 20% of all items are analyzed as the easy items, 46% of items are medium, and 34% of all items are difficult. The result of the study based on the analysis above shows that the question of SHS X can be indicated as not as good multiple choice question seen on the item facility, while for SHS Y and SHS Z can be indicated as good as multiple choice because they have low percentage of easy items and high percentage of medium items. The data summary of the item facility can be seen in detail in Graphic 1.



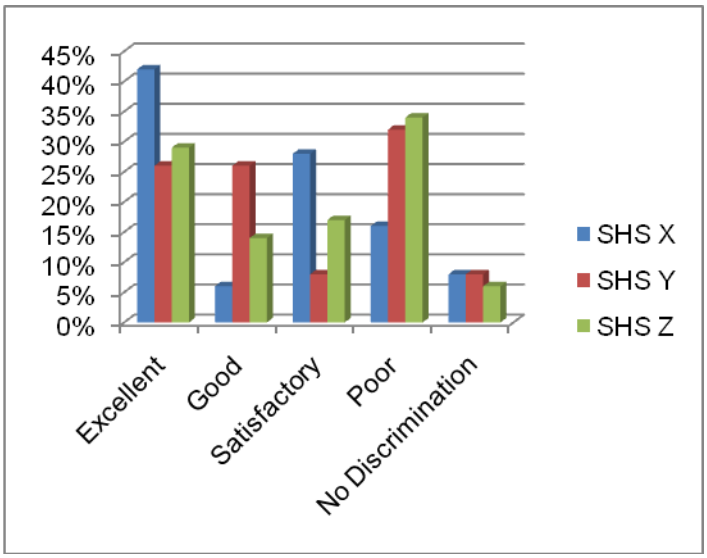
	Easy	Medium	Difficult
SHS X	50%	46%	4%
SHS Y	30%	50%	20%
SHS Z	20%	46%	34%

Graphic 1: Percentage of Item Facility

As a result in graphic 1, it can be seen that SHS X has a higher percentage for the easy items and a lower percentage for the difficult items than others. Whereas, SHS Z has a lower percentage of easy items and a higher percentage for difficult items than others. Accordingly, there are some possibilities why this kind of result happened, but it can depend on the students' ability of the items of the achievement tests. It is equally as stated by Bachman (1990: 19), that the students with higher ability are expected to have a higher probability of correct performances of the lower difficulty, and a lower probability of correct performances of greater difficulty, vice versa. Hence, it can be concluded that the students of SHS X have a higher ability than others.

2. Item Discrimination

As stated in the introduction section that the students' correct answers between high and low achiever are needed in this kind of analysis. Accordingly, the result of item discrimination is available in graphic 2 below, that is, SHS Y's achievement test is 42% of the test items belong to excellent items, 6% of items are good, 28% of all items indicated as satisfactory items, 16% are poor items, and 8% of items can not discriminate do not have the discrimination function. On the other hand, the excellent and good items of achievement test in SHS Y counted for each 26%, satisfactory items amounted 8%, poor items are in 32% of the test items, and 8% of the items can not discriminate the students' ability. Then, 29% of items are identified as the excellent items, 14% of all items are good items, 17% of test items are satisfactory items, 34% are indicated as poor items, and 6% are analyzed as items which do not have discrimination power.



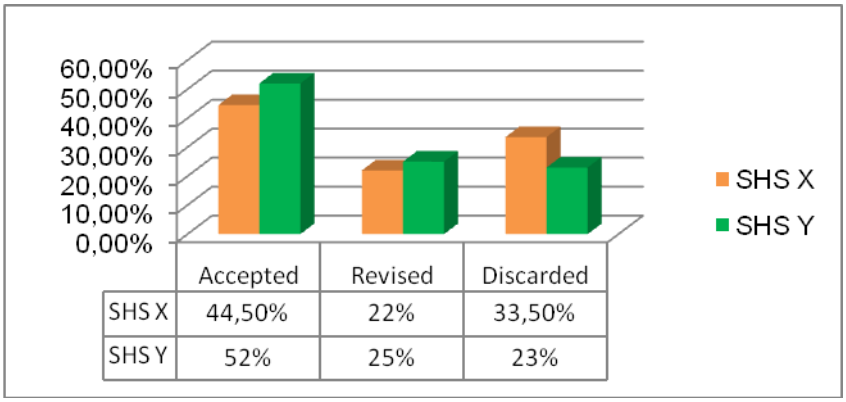
Graphic 2 The Percentage of Item Discrimination

Based on the analysis above, it can be said that the questions of achievement tests of three senior high schools have good discrimination items since the majority of the test items regarded to the satisfactory until excellent items.

It follows based on data above, almost half of items in SHS X regarded as excellent items, so it means that most of the items can discriminate the high and low ability students well, conversely, only 26% of 50 items on SHS Y are excellent. Again, SHS X has a lower percentage of poor items than others, therefore, only a few items in SHS X can not discriminate the students' ability, or it can be summarized that the items of SHS X are ideal. Even so, this study does not claim that the items of SHS Y and SHS Z are not ideal. It happens so since the items do not have the discriminating power as good as possible. Whereas the discriminating power is calculated from the students' response to the multiple-choice of the tests, thus, it depends on the students' ability. If most of the students can answer correctly, so the items neither do the function as excellent nor good items, instead, they do either as poor items or even no power items.

3. Distractor Efficiency

In this explanation, there are only two analysis of distractor efficiency, that is, the analysis of achievement tests of SHS X and Y. Incompletely data source of SHS Z is a reason of there are not distractor efficiency of it. Thus, there are 29% of 200 distractor items in the achievement test of SHS X are accepted, 37,5% are revised, and 33,5% are discarded. Yet, there are 42% of 200 distractors in SHS Y are accepted, 35% are revised, and 23% are discarded which can be seen in this following graphic 3.



	Accepted	Revised	Discarded
SHS X	44,50%	22%	33,50%
SHS Y	52%	25%	23%

Graphic 3: The percentage of Distractor Efficiency

The correct answers of the items are not counted as the distractors in this distractor analysis, likewise, J. D. Brown (1996: 71) explained that the distractors are those choices will be counted as an incorrect answer. The findings show there are almost half of distractors in SHS X are accepted, while more than half of distractors in SHS Y are accepted. Hence, it can be summarized that most of the items are good as those distractors success in diverting the students' answer. It does not run well to the discarded distractors which amounted almost quarter of distractors are discarded for both schools. Yet, it sounds good to happen since it gives information that many distractors do the function well.

4. Validity

The validity of the achievement tests have been analyzed by the researcher using the SPSS Statistics 22 which can be identified by looking at the index of Pearson Correlation higher than the index of r_{table} of total students, that is, 0,339 for the 34 students and 0,329 for the 36 students, then the test items are valid. The validity index of achievement test in SHS X counted as 76% items are valid and 24% items are invalid. In addition, there are 48% of items are valid, 52% of items are invalid for the achievement test of SHS Y. On the contrary, the achievement test on SHS Z identified 20% of all items are valid and 80% of items are invalid.

However, the findings of validity show there are still many invalid items from total items of three schools. In detail, SHS X has the highest percentage of validity, on the contrary, SHS Y has the lowest percentage of validity. For this reason, the SHS Y has the most invalid items than others. Nevertheless, this validity shows the ability of items in measuring the test items whether it measures what should be measured. In other words, it relates to the content of the test items, the material is given by the teacher, and the objectives of learning should be achieved. Thus, it can be summarized that the test items are still not appropriate with the real conditions.

5. Reliability

The result analysis of the reliability is based on the index of reliability. It must be higher than r_{table} of total students of the class, for instance, the r_{table} of 34 students is 0,339, then it can be concluded that the tests of SHS X and SHS Z are reliable, so did the test of SHS Y as the index of reliability is higher than r_{table} of students, 0,329. Therefore, the reliability index of the achievement test in SHS X is 0,905; SHS Y is 0,660 and SHS Z is 0,342.

For this reason, all of the achievement tests are reliable according to the data findings since all of the reliability indexes are higher than the r_{table} of each index. It shows all achievement tests have high consistency and it can be used for another test. Also, this consistency presents that the tests are dependable, means that the tests do the function to gather students' information in towards students' understanding.

The researcher elaborates and discusses the information collected in the previous research result based on the analysis of the multiple-choice of the achievement tests. This analysis of the achievement tests is taken from the theory of Brown in H. D. Brown (2004) and employed the Winstep-Rasch model.

CONCLUSION

As described in the previous findings and discussion, it can be drawn several conclusion: (1)The mean of item facility for SHS X, SHS Y, and SHS Z are 0,69; 0,55; and 0,44. The mean of those achievement tests indicates the tests are medium since the index of its mean is between 0,3 and 0,7. Thus, it is concluded the whole of those achievement tests fulfil the requirement as the good tests. (2)The mean of item discrimination for three schools are 0,326; 0,245; and 0,244. The index of SHS X examined as the good items, while the index of SHS Y and SHS Z analyzed as satisfactory items. It is concluded all the achievement tests are able to discriminate the higher and lower ability students. (3) Half of the distractors are accepted, so that, it does distract the students' answer. (4) One of three schools, SHS Z has valid index less than the total item of its test, thus most of the test items in SHS Z has not examined the students properly. Conversely, the index of three schools shows the tests are reliable and dependable. It means the whole of tests is consistent.

REFERENCES

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing* (Third Edit). Hong Kong: Oxford University Press.

Braun, H., Kanjee, A., Bettinger, E., & Kremer, M. (2006). Improving Education through Assessment, Innovation, and Evaluation. In *American Academy of Arts and Sciences* (p. 110). Cambridge: American Academy of Arts and Sciences. https://doi.org/10.3389/fnhum.2015.00046

Brown, H. D. (2004). *Language Assessment: Principles and Classroom Practices*. (V. L. Blanford, Ed.). New York: Pearson Education.

Brown, J. D. (1996). *Testing in Language Programs*. (D. Mosco, Ed.). New Jersey: Prentice Hall Regents.

Capkova, H., Kroupova, J., & Young, K. (2015). An Analysis of Gap Fill Items in Achievement Tests. *Elsevier Science Direct*, *192*, 547–553. https://doi.org/10.1016/j.sbspro.2015.06.087

Ciptaningrum, D. (2014). *An Item Analysis of English Summative Test on Difficulty Level and Discriminating Power*. Syarif HIdayatullah State Islamic University.

Cohen, L., Manion, L., & Morrison, K. (2007). *Research Methods in Education*. *Research Methods in Education* (Sixth Edit, Vol. 1). New York: Routledge Taylor & Francis. https://doi.org/10.1080/10572252.2010.502513

Fulcher, G. (2010). *Practical Language Testing* (First edit). London: Hodder Education.

Garvin, A. D., & Ebel, R. L. (1980). Essentials of Educational Measurement. *Educational Researcher*, *9*(9), 21. https://doi.org/10.2307/1175572

Grant, J., Allen, B., Anna, Edwards, A., & Henry. (2006). *Students Assessment Essentials Handbook*. *Academic Medicine* (Vol. 73). University of the West Indies. https://doi.org/10.1097/00001888-199809000-00035

Kumbakonam, U. R., & S, A. (2017). Role of A Teacher in English Language Teaching (ELT). *International Journal of Educational Science and Research (IJESR)*, *7*(February), 1–4.

Lebagi, D., Sumardi, S., & Sudjoko, S. (2017). The Quality of Teacher-Made Test in Efl Classroom At the Elementary School and Its Washback in the Learning. *Journal of English Education*, *2*(2), 97–104. https://doi.org/10.31327/jee.v2i2.289

Marie, S. M. J. A., & Edannur, S. (2015). Relevance of Item Analysis in Standardizing an Achievement Test in Teaching of Physical Science. *Journal of Educational Technology*, *12*(3), 30–36.