FENOMENA: *Jurnal Penelitian*
Volume 10, No. 1, 2018
e-issn 2615 – 4900; p-issn 2460 – 3902
DOI: http://doi.org/ 10.21093/fj.v10i2.1320

ANALYSIS ON ACHIEVEMENT TEST IN INTENSIVE ENGLISH PROGRAM OF IAIN SAMARINDA

Sari Agung Sucahyo
IAIN Samarinda
agungppsunm@yahoo.com

Widya Noviana Noor
IAIN Samarinda
noviana.widya@ymail.com

Abstract

As one of the tests, achievement test has to be qualified. A qualified test will be able to give the information about teaching correctly. If the achievement test is less qualified, the information related to students' sucesss to achieve the instructional objective will also be less qualified. It means the test has to meet the characteristics of a good test. In fact, there has not been any effort yet to identify the quality of the achievement test which is used in Intensive English program. It means the information of the test quality cannot be found yet. Therefore, researchers are interested in analyzing the quality of achievement test for students in Intensive English program of IAIN Samarinda. Design of this research belongs to Content Analysis. Subject of this research is English achievement tests and 28 to 30 students were involved in the process of try out. Data were collected through three steps. Data were analyzed based on validity, reliability, and item quality. Finding of the research reveals 60 % of the tests have a good construct validity justified by related theories. It was found 55% of the tests have a good content validity. Reliability coefficient of the first tests format is 0, 65 and the second tests format shows 0, 52. Calculation of item difficulty shows 68% of the test items were between 0,20 – 0,80. The estimation of item discrimination shows 73% of the test items were between 0,20 – 0,50. While calculation of distracter efficiency shows 65% of the distracters were effective to distract the test takers.

***Key-Words**: Achievement Test, English Intensive Program*

Abstrak

Sebagai salah satu jenis tes, tes prestasi haruslah berkualitas. Sebuah tes yang berkualitas akan memberikan informasi yang benar tentang pengajaran. Jika tes prestasinya kurang berkualitas, informasi yang terkait dengan keberhasilan siswa mencapai tujuan pembelajaran juga akan kurang berkualitas. Artinya, tes tersebut haruslah memenuhi karakteristik tes yang baik. Faktanya, belum ada upaya yang dilakukan untuk mengidentifikasi kualitas tes prestasi yang digunakan dalam program intensif bahasa Inggris Artinya informasi yang terkait dengan kualitas tes belum ditemukan. Oleh karena itu peneliti tertarik untuk menganalisa kualitas tes prestasi yang digunakan untuk mahasiswa pada program intensif bahasa Inggris IAIN Samarinda. Desain penilitian ini adalah Content Analysis. Subyek penelitian adalah tes prestasi dan 28 sampai dengan 30 mahasiswa dilibatkan dalam proses try out. Data dikumpulkan melalui 3 tahap sesuai aspeknya yaitu validitas, reliablitas, dan analisis butir soal. Hasil penelitian menunjukan 60% tes memiliki validitas konstruk yang baik didukung dengan teori terkait. Ditemukan 55% tes memiliki validitas isi yang baik. Koefisien reliabilitas dari tes format pertama adalah 0,65 dan yang kedua 0,52. Perhitungan tingkat kesulitan menunjukan 68% butir tes berada dalam rentang koefisien 0,20 – 0,80. Sedangkan perhitungan daya pembeda menunjukan 73% dari butir soal berada dalam rentang 0,20 – 0,50. Sementara perhitungan analisis pengecoh menunjukan 65% pengecoh dalam soal dipilih oleh peserta tes.

***Kata Kunci**: Tes Prestasi, Program Intensif Bahasa Inggris*

A. Introduction

State Institutes of Islamic Studies (IAIN) Samarinda has its own vision to be the excellent college in Kalimantan in 2025. The expected excellence refers to the quality of the institution to compete with other colleges in Kalimantan. One of the qualities is dealing with the academic outcome. It is expected that the alumnus of this college can cope with the competitive challenge in the future.

One of the academic qualities can be identified from the mastery of foreign language. English is an international language and plays an important role in supporting knowledge discovery. Referring to its important role, English is taught as a compulsory subject from junior high school up to college level. The mastery of English will be very beneficial to students in colleges to enhance their effort in exploring more knowledge and skills in their study process.

As the consequences, IAIN Samarinda provides English as one the courses taught in the curriculum. English is taught to students in the first and second semester and the name of the course is *Bahasa Inggris I* and

Bahasa Inggris II. In its implementation, the course is integrated with the *Pesantren* Program.

English course in IAIN Samarinda emphasizes the mastery of four English language skills; listening, speaking, reading, and writing. It is expected that students will be able to use those skills in communicative activities.This course is taught twice a week and each meeting is allocated to the two English skills. Therefore, four English skills can be equally taught during one semester. At the end of semester, final examination is done to measure whether or not students have succcesfully achieved the instructional obejective. In this case, the lecturers or instructors are assigned to construct the achievement test as the instrument of evaluation. Some lecturers were appointed and assigned in a team to accomplish the test and the current practice the test is constructed by each lecturer and it is administered in their own class.

As one of the tests, achievement test has to be qualified. A qualified test will be able to give the information about teaching correctly. Hughes explains that test has an effect to the instructional activities which is called backwash effect[1]. If the achievement test is less qualified, the information related to students' sucesss to achieve the instructional objective will also be less qualified. It means the test has to meet the characteristics of a good test.

In fact, there has not been any effort yet to identify the quality of the achievement test which is used in Intensive English program. It means the information of the test quality cannot be found yet. Based on the illustration above, it is necessary to identify quality of achievement test. Therefore, researchers are interested in analyzing the quality of achievement test for students in Intensive English program of IAIN Samarinda.

On the basis of the background, problem of the study is formulated in the following question, How is the quality of achievement test in Intensive English Program of IAIN Samarinda?

B. Review of Related Literature

A test can be called as good test, if it has criteria good test such as: validity, reliability, practicality, and item analysis.

Validity is the most important consideration in developing and evaluating measuring instruments. Historically, validity was defined as the extent to which an instrument measured what it claimed to measure. [2]

---

[1] Hughes, Arthur. *Testing for Language Teachers* (Cambridge: Cambridge University.2003) p.1

[2] Donald Ary, Lucy and Chris. *Introduction to Research Education.*Eight Edition.(New York: Holt,
Rinehart and Winston, 2006), p. 225

Validity is concerned with whether a test measures what it is intended to measure.[3] According to Franklen and Wallen and Ary et all in Budiharso Validity refers to the appropriateness, meaningfulness, and usefulness of the specific inferences researchers make based on the data they collect. Validation of instrument is the process of collecting evidence to support such inferences. The validity question is concerned with the extent to which an instrument measures with one thinks it is measuring.[4]

Validity divides into two areas including, logical validity, and empirical validity.[5] The logical validity provides logic or arguments by which an instrument is considered appropriate on the basis of logical or argumentative reasoning. The logical validity includes; content validity, criterion validity, and construct validity. Evidences used to support the logical validity do not necessarily use statistical analysis. In addition, empirical validity refers to validation of research instrument that use scores and statistical analysis as the basis of interpretation. The judgment of the appropriateness is based on the degree of the statistical calculation. The empirical validity includes criterion validity and predictive validity. The judgment of the appropriateness is based on the degree of statistical calculation. So, a valid instrument will provide correct conclusion and a researcher should draw based on the data collected.

The logical validity provides logic or arguments by which an instrument is considered appropriate on basis of logical or argumentative reasoning. The logical validity includes: (1) content validity, and (2) construct validity. Evidences used to support the logical validity do not necessarily use statistical analysis.

Content Validity Evidence, it is a test where if it has specific measurement and refers to the material which has been taught.[6] Content Validity is considered especially important for achieving his purposes as it is principally concerned with the extent to which the selection of test tasks is representative of the larger universe of tasks of which the test is assumed to be a sample.[7] Content validity is the appropriateness between the contents of teaching materials in curriculum or textbook and items of test. A good test should have good content validities. A good test is actually arranged based on teaching materials that have been taught before. If the items are appropriate with the teaching materials, the test has good content validities. Content validity is especially important in achievement testing.

---

[3] Cyril J.Wier.*Communicative Language Testing*. (America: Prentice Hall, 1990), p. 1

[4] Fraenkel JR and Wallen NE, *How to Design and Evaluate Research,* (New York: McGraw-Hill Inc, 1993), p. 139

[5] Norman E Grounlund, Op. Cit., p. 148

[6] Suharsimi Arikunto, Op.Cit., p. 67

[7] Cyril J. Weir, Op.Cit, p. 25

Building a test that has high content validity by (1) identifying the subject-matter topics and the learning outcomes to be measured, (2) preparing a set of specifications, which defines the sample of items to be used, and (3) constructing a test that closely fits the set of specifications. Content validity cannot be expressed in terms of a numerical index. Content validation is essentially and of necessity based on judgment, made separately for each situation. It involves a careful ad critical examination of the test items as they relate to the specified content area. One should determine whether the items in the test represent the course and objectives stated in the curriculum guides, syllabi, and texts.

Construct Validity Evidence, Construct- related to evidence of validity focuses on test scores as a measure of a psychological construct. To what extent do the test scores reflect the theory behind the psychological construct being measured? It is useful to assess individuals on certain psychological traits and abilities.[8] An Understanding of the concept of a psychological construct is prerequisite to understanding construct validity. A psychological construct is an attribute, proficiency, ability, or skill defined in psychological theories.[9] Fraenkel JR and Wallen NE identify three steps in obtaining construct validity: (1) the variable being measured is clearly defined, (2) hypotheses are formed in a particular situation, and (3) the hypotheses are tested both logically and empirically.

In addition, empirical validity refers to validation of research instrument that uses scores and statistical analysis[10]. Empirical validity refers to validation of research instrument that use scores and statistical analysis as the basis of interpretations. In addition, the empirical validity includes criterion validity and predictive validity.

Criterion validity refers to the relationship between scores obtained using the instrument and scores obtained using other instrument used as a criterion. The test is valid for purposes of predicting who will do well on the criterion. Criterion validity coefficients provide us with a useful basis for selecting and counseling students in various curriculum areas. But this is true only when a particular student is like the students who were in the sample on which the validly coefficient was determined.

The emphasis of this validity is on the criterion rather than the test contents. Criterion validity uses empirical techniques to study the relationship between scores on the instrument and the criterion. However, it must meet some criteria: relevance, reliable, and free from bias. Relevance

[8] Donald Ary e al, Op. Cit., p. 231

[9] James Dean Brown. *Testing in Language Programs,* (Prentice Hall Inc: Upper Saddle River, New Jersey,1996), p. 239

[10] Sri Esti wuryani D, *Psikologi Pendidikan*, (Jakarta: PT Gramedia Widiasarana Indonesia, 2006), p. 404

means the criterion must really represent behavior being tested. Reliable means the criterion must be a consistent measure of attribute over time or time-to-time.

There are two forms of criterion validity: predictive validly and concurrent validly.[11] Both are concerned with the empirical relationship between test scores and a criterion, but a distinction is made on the basis of the time when the criterion data are collected. Concurrent validity is the relationship between scores on a measure and criterion scores obtained at the same time.[12] Predictive validity is the relationship between scores on a measure and criterion scores available at a future time.

A key index in both forms of criterion validity is the correlation coefficient, indicating the degree of relationship of scores of two instruments. The coefficient correlation of two sets of scores is called as a validity coefficient. For example, a validity coefficient of 1.00 applied to the relationship between a set of aptitude-test scores (the predictor) and a set of achievement-test scores (the criterion) would indicate that each individual in the group had exactly the same relative standing on both measures, and would thereby provide a perfect prediction from the aptitude scores to the achievement scores.

The criteria of a good test includes at least the instrument should be reliable. According to Cryril J.Weir reliability is concerned with the extent to which we can depend on the test result. [13] According to Arikunto a test has high reliable if the test gives consistency the result of test.[14] Reliability is concerned with how consistently you are measuring whatever you are measuring.[15] According to Kerlinger in Dini Irawati summarizes reliability has similar meaning to dependability, stability, consistency, predictability and accuracy. [16] Test reliability is defined as the extent to which the results can be considered or stable.[17]

A test should give the same results every time it is used to measure. The results should be consistent and stable. In general, test reliable is defined as the extent to which the results can be considered consistent or stable. For example, when the teachers administered a placement test to their students on one occasion, the result of scores should be similar if they were to administer the same test again in different time.

---

[11] C.J Weir, *Understanding and Developing language Tests,* (America: Prentice Hall International English Language Teaching ,1993), p. 19

[12] Donald Ary et al, Op. Cit., p. 228

[13] Cyril. J.Weir. Op. Cit., p. 1

[14] Suharsimi Arikunto,…. 2002, Op. Cit., p. 86

[15] Donald Ary et al, Op. Cit., p.239

[16] Dini Irawati, M.Pd, Reliability of An Instrument: A Theoritical Review. (Samarinda: STAIN Samarinda, 2012), p.1

[17] Jame Dean Brown, Op. Cit., p. 192

The degree, to which a test is consistent, or reliable, can be estimated by calculating a reliable coefficient (r_{xx}). A reliability coefficient is like a correlation coefficient in that in can go as high as +1.0 for a perfectly reliable test. but reliability coefficient is also different from a cannot logically have less than no reliability. In cases when tester find negative values for errors; then if the calculations are all correct, they should round their negative result upward to 0 and accept that the result on the test had zero reliability.

Reliability coefficient or estimates as they are also called, can be interpreted as the percent systematic, or consistent, or reliable variance in the score on a test. for instance, if the scores on a test have a reliability coefficients of r_{xx} = 91, by moving to decimal two places to the right, the tester can say that the scores are 91 % consistent, or reliable, with 9 % measurement error (100%-91% = 9),or random variance. If r_{xx} = 40, the variance on the test is only 40 % systematic and 60 % measurement error.[18]

According to Donald Ary there are three categories of reliability coefficients used with norm referenced tests: (1) coefficients derived from correlating individual's scores on the same test administered on different occasions (test-retest coefficient), (2) coefficients derived from correlating individuals' scores on different sets of equivalent items (equivalent –forms coefficients), (3) coefficients based on the relationship among scores derived from individual items of subsets of items within a test (internal-consistency coefficients).[19] According to Brown that's language tester use three basic strategies to estimate the reliability as follows: the test-retest, equivalent-forms, and internal consistency strategies. Commonly, the technique uses statistical formula applicable to analyze as follows: Product Moment, Spearman-Brown, KR-20, and KR-21.

Test-Retest reliability is the one most appropriate for estimating the stability of a test over time. According to Donald Ary an obvious way to estimate the reliability of a test is to administer it to the same group of individuals on two occasions and correlate the two sets of scores. The correlation coefficient obtained by this procedure is called a test-retest reliability coefficient. The first step in this strategy is to administer whatever test is involved two times to group's students. The testing sessions should be far enough apart so that students are not likely to remember the items on the test. Once the tests are administered twice and the pairs of scores for each student are lined up in two columns, simply calculated a Pearson Product-moment correlation coefficient between two sets scores. The correlation coefficients will provide a conservative estimate (that is , a low estimate, or underestimate) of the reliability of the test over

[18] James Dean Brown, Op. Cit., p. 193
[19] Donald Ary et Al, Op. Cit., p. 242

time.[20] This reliability variance on the test. However, situations do occur in which the test-retest strategy is the most logical and practical alternative for estimating reliability.

Equivalent–forms reliability (sometimes called parallel-forms reliability) is similar to test-retest reliability. However, instead of administering the same test twice, the tester administers two different but equivalent tests (for example, forms A and B) to a single group of students. The tester calculates correlation coefficients between the two sets of scores on the two forms. The resulting equivalent-forms reliability coefficient can be directly interpreted as the percent of reliable, or consistent, variance on either form of the test. However, it is strategy provide an estimate of the consistency of scores across forms rather than over time, as the case with test-retest reliability.

Other reliability strategies are designed to determine whether all the items in a test are measuring the same thing. These are called the internal-consistency strategies and require only a single administration of one form of a test.

Split –Half Reliability is the easiest internal – consistency strategy to understand conceptually is called the split-half method. This approach is very similar to the equivalent-forms technique expect that, in this case, the equivalent forms" are created from the single test being analyzed by dividing into two equal parts. The test is usually split on the basis of odd- and even-numbered items. The odd and even numbered items on the test are scored separately as though they were two different forms. A correlation coefficient is then calculated for the two sets of scores.[21] The most common procedure, however, is to correlate the scores on the odd-numbered items (*X*) of the test with the scores on the even-numbered items (*Y*). Usually, this calculated represents the degree of reliability for only half of the test-either half, but still just half of the test.

Kuder-Richardson (K-R) Formula. Kuder and Richardson developed procedures that have been widely used to determine homogeneity or internal consistency. Probably the best known index of homogeneity is the Kuder-Ricahrdson formula 20 (K-R 20), which is based on the proportion of correct and incorrect response to each of the items on a test and variance of the total scores:

$$r_{xx} = \frac{K}{K-1}\left(\frac{s_{x-\sum pq}^2}{s_x^2}\right)$$

KR-20 where: r_{xx} = reliability of the whole test

[20] Brown, Op. Cit., p. 193

[21] JD Brown, Op. Cit., p. 194

K = number of items on the test
S_x^2 = variance of scores on the total test (squared standard deviation)
P = proportion of correct responses on a single item
q = proportion of incorrect responses on the same item[22]

Another formula is Kuder-Richardson 21, is computationally simpler but requires the assumption is often unrealistic:

$$r_{xx} = \frac{Ks_x^2 - \bar{X}(K-\bar{X})}{s_x^2 (K-1)}$$

Where: r_{xx} = reliability of the whole test
K = number of items in the test
S_x^2 = variance of the scores
$\bar{X}$ = mean of the scores

Reliability that uses a single administration of a single form is based on the consistency of responses to all items in the test. Kuder-Richardson (K-R) formula measure the internal consistency, estimated from a single administration of a test through a study of score variance. Instead of comparing scores on different administrations of a test, it is far easier to estimate reliability by comparing scores on the test's items, considering each item as a test in itself. If the items show a high degree of estimate, the test is an accurate or consistent measure.[23]

Besides validity and reliability as main characteristic, there is another characteristic that supporting a good quality of test, namely the third characteristics of good tests are practicality or usability in the preparation of a new test.

According to Arikunto, there are some aspects the test has high practicability, when the test has practical characteristic, easy to administering. That practical test namely: 1) Easy to construction, 2) easy to administration, 3) easy to scoring.[24]

According to Djiwandono, there are some aspect that need to discover, such us; practicality is not demand of using difficult facility (it means test with practicality by using facility that usually used to daily teaching and learning activity), as the teacher must keep in mind a number of very practical considerations which involves economy, ease of administration, scoring and interpretation of result. Economy means the

[22] Donald Ary et al, Op. Cit., p. 245

[23] Tuckman, *Measuring Educational Outcomes: Fundamental of Testing, (New York:Harcourt Brace Jovanovich, Inc, 1975),* p. 256

[24] Suharsimi Arikunto, …2002, Op. Cit., p.62

test is not costly. The teachers must take into account the cost per copy, how many scores will be needed, (for the more personnel who must be involved in giving and scoring a test, the more costly the process becomes). How long the administering and scoring of it will take, choosing a short test rather than longer one. Easy to administration and scoring mean that the test administrator can perform his task quickly and efficiently. We must also consider the ease with which the test can be administered.

Besides having a good criteria, the other characteristics of the test that more important and specific is the quality of the test items. To know the quality of the test items, teachers should use a method called item analysis. There are several meanings of what item analysis. According to Anthony J Nitko, in his book, he stated that: Item analysis refers to the process of collecting, summarizing, and using information about individual test items especially information about pupils' response to items".[25]

Item analysis is an important and necessary step in the preparation of good multiple choice test. Because of this fact; it is suggested that every classroom teacher who uses multiple choice test data should know something of item analysis. How it is and what it means.[26]

For the teacher made test, the followings are the important uses of item analysis: determining whether an item functions as teacher intended, feed back to students about their performance and as a basis for class discussion, feedback about pupil difficulties, and area for curriculum improvement, revising the item and improving item writing skill. Item analysis usually provides two kinds of information on items[27]:

a) Item difficulty (p)

Item difficulty is statement about how easy or difficulty an item for the test taker. Item difficulty of a test, reveal the level of difficulty of a test. The item difficulty can state whether a test is very difficult, difficult moderate, easy or very easy. An item which can be answered by most or even all of test takers is classified into easy or even very easy item. In the contrary, an item which cannot be answered by most or even all of test takers is classified into difficult or even very difficult item. An ideal item is an item which is not too easy or too difficult.

The steps to measure the item difficulty are as follows:

1. Identify the answers of the subject toward all items.
2. Put the answers in the tabulation.

---

[25] Anthony J. Nitko, *Educational Test and Measurement an Introduction*, (New York:Harcourt Brace Jovanich inch., 1983), p.284

[26] Jhon W. Oller, *Language Test at School* , (London: Longman group, 1979), p. 245

[27] H.G Widdowson, *Language Testing* (Oxford: University Press., 2000), p.60

3. Add up the number of the subject who correctly who took a particular item
4. Divide that sum by the total number of subjects who took the test.

b) Item discrimination (D)

Item discrimination (D) indicates the degree to which an item separates the students who performed well form from those who performed poorly. According to Brown, these two groups are sometimes referred to as high and low-scores of upper and lower-proficiency students. Analysis of item discrimination addresses a different target: consistency of performance by candidate's across items. The usual method for calculating item discrimination involves comparing performance on each item by different groups of test takers: those who have done relatively poorly. For example, as items get harder, we would expect those who do best on the vest overall to be ones who in the main get the right. Poor item discrimination indices are signal that an item deserves revision. If there are a lot of items with problems of discrimination, the information coming out of the test is confusing, as it means that some items are suggesting certain candidates that relatively better, while order individuals are better, no clear picture of the candidates. Ability emerges from the test (The scores, in other words, are misleading and not reliable indicators of the underlying abilities of the candidates) such a test will need considerable revision.

C. Review of Previous Studies

The first study was conducted by Zubair Haider, Farah Latif, Samina Akhtar, and Maria Musthaq (2015). The title of their research is Evaluation of English achievement test: A comparison between high and low achievers amongst selected elementary school students of Pakistan. The result of this research are (1) For the English achievement test, item difficulty was computed for each item, where the value of item difficulty was greater than 80% and less than 20%, those items were rejected because they are very easy and very difficult item. (2) Items having discrimination index = 0.20 or less were rejected, because they were unable to discriminate between HA and LA on the basis of the set criterion. (3) On the basis of mean performance in test score, it is observed that male mean performance is better than female students for 8th grade in selected elementary schools. (4) The combined mean and combined standard deviation of male group was greater than female group in the test,the calculated value of Z-test 1.80 is also smaller than table value 1.96, the difference among male and female mean performance is statistically significant, which means male students are better performer than female students at 8th grade in selected elementary schools in Pakistan. (5) Calculated value of reliability coefficient

was more than 50 of the three methods, split-half (0.74), KR20 (0.78), and KR21 (0.70), which means the test was concerned to be reliable to great extent.

The second study was conducted by Halka Capkova, Jarmila Kroupova, and Katerina Young (2014). The title of this research is An Analysis of Gap Fill Items in Achievement Tests. Finding of their research proved that there are some weak items which do not work properly and therefore should be replaced or changed. All aspects taken into account revealed that gaps 3 and 10 do not correspond to the range of both FI, DI and do not serve as the intended or an additional distractor. A similar situation appeared in case of gaps 4 and 9 which showed some satisfactory values but in two other aspects they did not work.

The last research was conducted by Gemma R Pascual (2016). The title of this research is Analysis of the English Achievement Test for EFL Learners in Northern Philippines. This research showed that he test was reliable in terms of internal consistency and reliability. The test contains items that could satisfy the purpose of the test. The reliability of the test was higher (0.898) as corrected by the Spearman formula. Findings imply that the English Achievement Test prepared and administered among the CSU sophomore students is valid and reliable. For the overall difficulty and discrimination of the test, the mean score of the 99-item test was 40.16. The test was considered difficult compared to the ideal mean of 61.87. The test had a discrimination index of 0.37 as shown in the coefficient of variation. The test does not discriminate well.

D. Methodlogy

The research uses Content Analysis as the design. Donald Ary et al defines "Content analysis focuses on analyzing and interpreting recorded material within its own context. The material may be public records, textbooks, letters, films, tapes, themes, reports, and so on. [28]

In this study, the researchers analyzed the quality of achievement test in intensive English Program of IAIN Samarinda. The subject of the try out is achievement test in intensive English program of IAIN Samarinda. The tests were taken from the last one year. 28 to 30 students who currently join intensive English program were involved in the process of test try out.

The instruments were used in this study is documentation. In this case the documents are files of achievement test and also the text books used in the instructional process.

[28] Donald Ary, Lucy C, Asghar .R. Introduction to Research in Education sixth edition, (Northem Illinois University,2002), p. 24

To collect data, three steps were taken by Asking for the documments of the achievement test from the management of intensive English program, Classifying the tests into their category based on the four English skills, and Trying out the tests to the real target students.

Data of this research were analyzed based on some criteria of a good test as follows: (1) Validity. In this study, the researcher uses qualitative approach to shows the result of validity that consists of content validity and constructs validity. Analysis content validity shows the test based on the teaching materials or indicator on syllabus. Then, analyses construct shows the test based on the theory. Theory means language skills and language components theory; (2)Reliability. In this study, the researcher uses quantitative approach to shows the result of reliability by using Split-half methods technique. According to Brown that's the easiest internal – consistency strategy to understand conceptually is called the split-half method[29]. Split-half method is measure of internal consistency of a test, because only a single administration between frequency on odd-numbered and even-numbered items of a single test. In this study, the researcher chooses one of way to find out reliability by using Split-half method. In Split-half method, that will find test that fulfill three criteria item analysis, namely; r item, difficulty index, item discrimination. Then distribute item in two groups, meanly odd number and even number. The calculated of coefficient correlation with correlation product moment formula, where X as odd score number and Y as even score number. To transform the split-half correlation into an appropriate reliability estimate for the entire test, the Spearman-Brown Prophecy formula is employed.

In the test which is not represented by items or numbers like writing test, inter-rater reliability method was applied. Item analysis was done to estimate item difficulty, item discrimination, and also distracter efficiency of the test

E. Research Finding

Finding of this reasearch was obtained from two processes. The first one is dealing with logical interpretation about quality of the test and the second one was from the empirical evidence by conducting the try out to the test takers.

1. Validity

Construct validiy of the test was analyzed by verifying description of the test with the underlying theories of language skills.

In the aspect of construct validity, underlying theories about the skills or area to be measured by the tests were used to find out the

[29] JD Brown, Op.Cit., p. 194

alignment between what the theories and the tests. It was found that 60 % of the tests have identical construction justified by related theories. 40% of the tests do not have a good construction. It means the 40% need to be justified by related theories.

Related to content validity, the estimation was done by comparing instructional content of English Course with the tests coverage. It was found 55% of the tests have a good content coverage and 45% of the tests do not have a good content coverage. Therefore, 45% of the tests need to cover more items in the instructional content of English Course.

2. Reliability

Concerning with reliability estimation, the tests were divided by two formats. The first format is tests which provide some items to answer. The second format is the tests which require students or test takers to produce utterances both in spoken and also written form as it was found in speaking and writing section. Reliability of the test was calculated based on two methods as follows:

1. Split-Half Method
   Reliability coefficient of the first tests format is 0, 65
2. Inter-Rater Method
   Reliability coefficient of the second tests format shows 0, 52.
3. Item Quality Analysis
   Item analysis was done in order to find out quality of the items which cover item difficulty, item discrimination, and distracter efficiency

1. Item Difficulty
   Calculation of item difficulty shows 68% of the test items were between 0,20 – 0,80.
2. Item Discrimination
   The estimation of item discrimination shows 73% of the test items were between 0,20 – 0,50.
3. Distracter Analysis
   calculation of distracter efficiency shows 65% of the distracters were effective to distract the test takers.

F. Discussion

Firstly, related to validity of the test. Finding of the research reveals 60 % of the tests analyzed have a good construct validity. Meanwhile, 40% of the tests do not have a good constrtuct validity. The tests do not really measure language skills that are supposed to be measured. The test require students or test tasker to choose the options provided by crossing. The tests should let the test takers to perform their language skills as authentic as possible. The choice of format may affect construct validity of the test. It is not a good idea to measure students' achievement in speaking and writing

by appling multiple choices test formats. And then referring to content validity of the test, research finding shows 45% of the tests analyzed were still less qualified. It means the tests do not properly represent both language skills and also language coverage which is stated in the instructional content. Speaking is a language skill which is rarely measured by the instructors in their achievement test. An achievement test has to be able to represent instructional content so the test can measure what is intended to measure in the instructional program.

Secondly, concerning with reliability of the test. Based on the try out process using split half and inter rater methods, the test developed by the instructors shows an acceptable coeficcient range both the first and also the second type of test format. It means the tests generally have a good reliability from moderate to high reliability coeficcient category. A good reliability of the test may help to inform data preciseness of students ability being measured.

And then in relation to item difficulty, it was found that 68% of the test have ideal level of difficulty. The rest of items have to be dropped or revised because they are both too easy and also too difficult for test takers to do. A good achievement test items should not be either too easy or too difficult.

Furthermore, referring to item discrimination, finding of this research indicates items of the tests in the accpetable coeficcient range. It means items of the test generally have the abilty to distinguish test takers based on their ability. So it is clear to find out and identify low proficient and how proficient test takers.

Finally, it has something to do with distracter efficiency. Finding of this research has clearly shown that 65% of distracters are efffective to distract test takers when they did the tests. But 35% of distracters were still not effective . When multiple choices test formats applied in the test, it is very important to create distracters as effective as possible so that if students can answer the question it happens because they really know the answer and not because of gambling.

G. Conclusion

This research is intended to identify quality of achievement test for students in Intensive English program of IAIN Samarinda. On the basis of research finding and discussion, conclusions are formulated as follows, In general, the tests have a good quality in validity, reliability, and item quality. It was found some parts of the tests which is dealing with validity and item quality were still less qualified.

Based on the conclusion, related suggestions are recomended as follows, Lecturers or instructors need to maintain their efforts to have good

test items. They are required to pay attention to procedure of developing the test. Lecturers or instructors have to write test blue print or test specification before developing the test in order to make sure every item within the test on the right track based on the construction and also content as well. It is also recommended to have a peer review before the test is administered. For further researches, it is suggested to extent the sample of the test in order to display and inform more finding about test quality.

REFERENCES

Anthony J. Nitko, *Educational Test and Measurement an Introduction*, (New York: Harcourt Brace Jovanich inch). 1983

Budiharso, T. 2005. Evaluasi Berbasis Kelas dalam Pembelajaran Bahasa Kedua. Jakarta: Lekdis.

Cyril J.Wier.*Communicative Language Testing*. (America: Prentice Hall,),1990

Donald Ary, Lucy and Chris,. *Introduction to Research Education.*Eight Edition. 2006

Fraenkel JR and Wallen NE), *How to Design and Evaluate Research,* (New York : McGraw-Hill Inc. 1993

Gemma R. Pascual. Analysis of The English Achievement Test for ESL Learners in Northern Philippines*, Internasional Journal of Avanced Research in Management and Social Sciences*. Vol. 5 No. 12 December 2016

Halka Capkova, Jarmila Kroupova, and Katerina Young. 2014. An Analysis of Gap Fill Items in Achievement Tests. *Procedia Social and Behavioral Sciences*. Dubai-United Arab Emirates: December 11-13, 2014. Hal. 553

H Douglas Brown , *Teaching by principles an Interactive Approach to Language Pedadogy*, (San Fransisco University: Longman Second Edition,).(New York: Holt, Rinehart and Winston) .2001

Hughes A., Testing for Language Teachers. Cambridge. Cambridge.2003

James Dean Brown. *Testing in Language Programs,* (Prentice Hall Inc: Upper Saddle River, New Jersey,) .1996

M. Soernadi Djiwandono, *Tes bahasa dalam pengajaran,* (Bandung: ITB,). 1996

Norman E. Gronlund, *Constructing Achievement Tests* Third Edition, (New York: Macmillan Publishing.co. 1985

Norman E. Groundlund, *Measurement and Evaluation in Teaching*, (New York:: Macmillan publishing Co., Inc.,). 1981

Oxford University, *Oxford Learners' Pocket Dictionary*, (New York: Oxford Unversity Press,). 2008

Sri Esti wuryani D, *Psikologi Pendidikan*, (Jakarta: PT Gramedia Widiasarana Indonesia,) 2006

Suharsimi Arikunto, *Dasar- Dasar Evalusi Pendidikan: Edisi Revision,* Jakarta:Bumi Aksara. 2001

Tuckman, *Measuring Educational Outcomes: Fundamental of Testing,* (*New York:Harcourt Brace Jovanovich, Inc* University Press.*1975*

Wilmar Tinambunan, *Evaluation of students achievement,* (Jakarta: Depdikbud. 1998.

Zubair Haider, Farah Latif, Samina Akhtar, and Maria Musthaq. Evaluation of English Achievement test: A comparison between high and low achivers amongst selected elementary school students of Pakistan, *Internasional Scholers Journals*. Vol. 3 (3), pp. 141-148, May, 2015