

ANALYSIS OF SUMMATIVE TESTS FOR ENGLISH

By
Rusma Setiyana*

University of Syiah Kuala, Banda Aceh

ABSTRACT

This study aimed to analyze the quality of summative tests for English at MAN Boarding School Meulaboh I in terms of validity, reliability, difficulty index, discrimination index, and the effectiveness of distractors. Content analysis was employed in this study. Two techniques were carried out to collect the data, namely a checklist and document analysis. The data from the checklist was analyzed using statistical procedures and the data from the document analysis was analyzed using Anates software version 4. The results showed that the validity of the English summative tests at MAN Meulaboh I was on average either sufficient or poor since the percentages were below 72%. Secondly, the tests had a high and consistent degree of reliability. The index of difficulty was above 70%. Thirdly, 60% of the difficulty index in the test of the first grade, 48% in the second grade, and 8% in the third grade test were accepted. Fourthly, more than half of the discrimination index was good. In detail, good in the discrimination index of the test was 76% in the first grade, 56% in the second grade and 72% in the third grade. Finally, the effectiveness of distractors in the English summative test in the first grade was 53%, in the second grade was 67% and in the third grade was 50%.

Key words: *Summative Test, Item Analysis.*

INTRODUCTION

Assessment plays an integral role in the teaching-learning process. It is employed to identify how students are learning and understanding lessons taught at school. Many kinds of assessment are used by

* Corresponding author: thiev_a@yahoo.com

teachers in the classroom as teachers must always be concerned with whether they have reached the goals for their teaching efforts. Tests as part of assessments are required to provide information about the achievements of learners, and are usually carried out at the end of a semester/course; these are traditionally called summative tests. Specifically, they aim to determine whether students should be allowed to advance to the next stage to deal with new, more advanced, objectives of learning. Without any tests, a teacher will have difficulty in providing proof of the quality of his students.

To construct a good test that is fair is not easy to do. A teacher needs to work hard. Brown (2004) sets out several stages of test construction which consist of determining test objectives, drawing up test specifications, devising test tasks, scoring, grading, and giving feedback. To produce a better one, a teacher must follow the available syllabus and deal with many references related to the rules on how the test items should be made. As a consequence, the teacher is not allowed to make a test based on his own desires without referring to the syllabus.

A teacher has to consider the quality of a test while creating an effective test. In this case, a test must possess two important characteristics, namely validity and reliability. This means that a test has to be well constructed by considering the validity and reliability of the test in order to provide an accurate measure of the test-taker's ability within a particular domain and to ensure the test is appropriate to what was taught in the class. Furthermore, a teacher is supposed to be familiar with item analysis which involves three parts, namely the index of difficulty, the discrimination index, and the effectiveness of distractors in order to increase the quality of the teacher made test. However, many teachers lack the basic knowledge about these and lack awareness of the importance of analyzing the test items they make.

This study was done to analyze the quality of a summative test for English taken by the students at MAN Boarding School 1 in Meulaboh, Aceh Barat in terms of validity, reliability, difficulty index, discrimination index, and the effectiveness of distractors.

LITERATURE REVIEW

Assessment

Assessment represents a vital element in language instruction. Popham (2003) describes assessment as a measure or activity to obtain

valid information about a student's knowledge, skills and/or ability. Harlen (2005) emphasizes that all assessments must involve ways to collect the evidence, ways to interpret it and ways to communicate it to the test takers in order to make a decision to have relevant evidence for a certain goal. As a result, they, the students, will become more motivated, and enhance their performance for their language instruction.

Assessment is a very important part of the teacher's job. Teachers must prepare a variety of procedures and tools to meet the needs of all learners and all situations and to find out to what extent their methodology, materials and tests are effective.

The Concept of a Test

Roszkowski and Spreat (2011) define a test as any systematic procedure for obtaining information about persons, objects, or situations. Tests have many purposes for teachers, students, and instruction programs. In particular, Basanta (2012) has written that tests are being used to:

- (a) know students' progress,
- (b) show the teacher how successful their teaching has been,
- (c) know students' strengths and weaknesses and to
- (d) evaluate the effectiveness of the program, the course book, and how the teaching-learning needs were met.

Types of Tests

Brown (2004) points out that a test conducted during the teaching-learning process is known as a formative test. This type of test includes all kinds of informal assessments done in the classroom. On the other hand, according to Harlen (2005) a summative test refers to the process by which teachers collect data systematically to make an inference about the state of a student's learning.

Thomas et al. (2004) claims that three types of test most commonly used by teachers are (i) standardized tests, (ii) commercially prepared assessments, and (iii) teacher made tests. As can be seen in the website of Asia E-University (2009), teacher made tests are occasionally more effective than standardized tests. Unlike standardized test, Walker and Schmidt (2004) mention that teacher made tests are designed not to compare or to provide students' ranking. Besides, tests also are divided into direct tests and indirect tests. According to Brown (2004), direct tests include tests that require the test taker to perform the target task.

Indirect tests mean that the test takers are not involved in actually performing the task itself but rather a task that is related in other ways.

Characteristics of a Good Test

Validity

Liao (2004) defines test validity as the extent to which inferences made from assessment results are appropriate to test scores. Whereas according to Hughes (2003) validity refers to whether the test measures accurately what it is intended to measure.

Hughes (ibid) divides validity into four parts: face validity, content validity, construct validity, and criteria validity. Fulcher and Davidson (2007) define content validity as whether the test refers to a representative sample from the domain that is to be tested. McCowan and McCowan (1999) mention that content analysis contains the degree to which a test matches a curriculum and accurately measures the specific training objectives on which a program is based. In short, content analysis involves whether the test given represents the whole of the material taught by the teacher and whether the test is appropriate within the requirements of the intended curriculum.

Another aspect of validity is face validity. Harrison (1991) says that the way to discover face validity is by asking the opinion of the students and the teachers concerned regarding the validity of the test formally or informally. Next, Okunya (2014) defines construct validity as an effort to ensure that the test is essentially measuring the intended attribute and not other extraneous ones. As a result, the test cannot measure the test takers' ability accurately if it does not measure what the teachers intended it to measure.

The criteria of validity are separated into concurrent validity and predictive validity. Fulcher and Davidson (2007) say that a test has concurrent validity if there is a test score which can be correlated to the criterion at about the same time. Whereas Hughes (2003) defines predictive validity as the degree to which a test can predict candidates' future performance. This means that predictive validity does not engage any available external criteria when the test is developed.

Apart from that, item validity can be calculated statistically by using a coefficient of correlation from a student's own scores. Fulcher and Davidson (2013) call it the validity coefficient of the test, which is the coefficient of correlation between the score of the test and the score of the criterion. It can be measured by using the point bi-serial formula which Brown (2004) offers as follows:

$$r_{pbi} = \frac{M_p - M_q}{S_t} \sqrt{pq}$$

Where:

rpbi = point bi-serial correlation coefficient

Mp = whole-test mean for students answering item correctly

Mq = whole-test mean for students answering item incorrectly

St = standard deviation for whole test

p = proportion of students answering correctly

q = proportion of students answering incorrectly

Reliability

Roegier (2014) refers reliability to the dependability of test scores. That is the degree to which a test-maker can expect relatively similar test scores from individuals across testing situations on the same, or parallel, testing instruments. Khalifa and Weir (2009) divide reliability into three categories. They are alternate-form coefficient, test-retest coefficient, and internal consistency coefficient. For internal consistency, KR20 is one formula suggested by Heaton (2000) to estimate reliability. The formula is set out below:

$$r_{11} = \frac{N}{N-1} \left[1 - \frac{m(N-m)}{N_x^2} \right]$$

Where N refers to the number of test items, m is the proportion of people passing the item, $(N-1)$ is the proportion of people failing the item, and x represents variance.

Difficulty Index

According to Asaad and Hailaya (2005) the difficulty index refers to the proportion of the upper group and of the lower group who answered an item correctly. A formula suggested by Brown (2004) to find the index of difficulty is set out below:

$$Dy = \frac{\text{\# of Ss answering the item correctly} \times 100}{\text{Total \# of Ss responding to that item}}$$

Discrimination Index

Brown (ibid) also defines the discrimination index as whether an item is able to differentiate between a low and an upper group of test-

takers. This means that a test with poor discriminating power will not give a reliable interpretation of the students' real ability. To calculate the discrimination index, Heaton (2000) offers the following formula:

$$Dn = \frac{\text{Correct } U - \text{Correct } L}{n}$$

Where Dn is the discrimination index, n refers to the number of test takers; U represents upper half and L lower half. The index of discrimination is on a range 1.00 to 0.00.

Distractors

Distractors play a vital role to ensure the quality of multiple choice tests. As Brown (ibid) has stated using distractors is one of the ways to measure the effectiveness of multiple choice tests. The distractors work as a lure option to make a test taker choose the wrong choice. Burton, et al. (1991) say that it is indeed the aim of distractors to appear as a plausible option for those students who have not achieved the objective being measured. Sunarya (2003) offers a formula to estimate the effectiveness of distractors.

$$IPc = \frac{nPc}{(N - nB) / (Alt - 1)} \times 100\%$$

Where:

IPc = effectiveness of distractors

nPc = the proportion of students who take the distractors

N = total students who follow the test

nB = the proportion of students who answer the item correctly

Alt = the total of choices

RESEARCH METHOD

The design of this study followed the format for content analysis research using a quantitative method. The data collected was analyzed using statistical software Anates Version 4.

Object of Study

The object of this research was English summative test items taken by all students in each grade. The student population totaled 179

students; 35 girls and 30 boys in the first grade (year 10), 36 girls and 27 boys in the second grade, and 33 girls and 18 boys in the third grade.

Data Collection Technique

Two techniques were carried out to collect the data, namely a checklist and document analysis. The checklist was one of the instruments used to analyze the test validity. In this case, a test specification was used to analyze content validity and construct the validity of the summative test for English. Meanwhile, the document analysis in this study used the syllabus, the summative test for English, and the students' answer sheets.

Data Analysis

To analyze content validity, the test items were compared to the demands of the syllabus content including the standard competency and basic competency by using test specifications. This would be the proof whether the teacher-made test items were appropriate to what the students had learned in their classrooms. Meanwhile, construct validity can be analyzed by using the summative test for English.

Validity coefficient, reliability, discrimination index, difficulty index, and the effectiveness of distractors were analyzed using Anates Version 4. The Anates software was used to analyze the raw data to get a result. The results from the Anates version 4 analyses were then compared to each criterion via: for reliability, the discrimination index, the difficulty index, and the effectiveness of the distractors.

FINDINGS AND DISCUSSIONS

Findings

The content validity of the summative test for English at MAN Meulaboh I at the end of the academic year 2014/2015 was varied. Related to the results from the analysis of basic competency, it indicated that the first grade test had sufficient validity (62.5%), the second grade test had poor validity (50%) and the third grade test had very poor validity (25%). In detail, there were 13 basic competencies which were ignored by the English teachers at MAN Meulaboh I. They were basic competency points 4.1, 4.2, and 6.1 in the first grade, points 3.2, 4.1, 4.2, and 6.1 in the second grade, and points 3.1, 3.2, 4.1, 4.2, 5.1, and 6.1 in the third grade. Moreover, it was found that there were 16 items which did not match with the English syllabus; 8 items

(numbers 14, 15, 16, 17, 18, 23, 24, 25) in the first grade test, 3 items (numbers 18, 24, 25) in the second grade test, and 5 items (numbers 6, 7, 8, 9, 10) in the third grade test.

Whereas in the context of construct validity, two of the summative tests for English had poor construct validity. In detail, the construct validity of the test for the first grade was poor (48%), that for the second grade was sufficient (72%) and that for the third grade was poor (50%). Otherwise, there were some items which had low construct validity because they included indirect tests. These were items numbers 1, 2, 3, 4, 5, 6, 19, 20, 21, 22 in the first grade test, items numbers 15, 16, 17, 18, 19, 20, 24, 25 in the second grade test, and items numbers 21, 22, 23, 24, and 25 in the third grade test.

The findings also show that in terms of the validity coefficient 60% of the items in the summative test for English for the first grade were valid, via: items number 1, 2, 5, 6, 10, 11, 14, 17, 18, 19, 21, 22, 23, 24, and 25. The result for the validity coefficient for the third grade test was similar via: items number 2, 7, 8, 11, 12, 13, 14, 15, 16, 17, 19, 20, 22, 23, and 24 were valid. However, only 44% of the test items for the English summative test for the second grade were valid. These were items number 3, 4, 5, 7, 8, 16, 17, 19, 22, 23, and 24.

Thus the summative tests for English at MAN Meulaboh 1 were reliable. This was proven with the high reliability index for the summative test for English in each grade. In detail, the index of reliability for the English summative tests in both the first and the second grade was 0.70, whereas the index of reliability for the summative test for English for the third grade was 0.88. Subsequently, the results for the difficulty index for the Summative Tests for English at MAN Meulaboh I are in Table 1.

Table 1. Results for test items for index of difficulty.

Index of Difficulty	Interpretation	I	II	III
0 – 15%	Very difficult	16%	0%	0%
16% - 30%	Difficult	8%	0%	4%
31% - 70%	Moderate	60%	48%	8%
71 % - 85%	Easy	12%	24%	60%
86% - 100%	Very easy	4%	28%	28%

It can be seen from Table 1 that the summative test for English for the first grade had a higher proportion of moderate difficulty index items (60%) than the other grades. Otherwise, the second grade had a

moderate to very easy difficulty index and the third grade predominantly had an easy difficulty index (60%).

Besides that the results for the discrimination index from the summative tests for English were as shown in Table 2 below:

Table 2. Results for index of discrimination.

Index of Discrimination	Interpretation	I	II	III
50% - 100%	Excellent	36%	20%	32%
30% - 49%	Good	40%	36%	40%
20% - 29%	Fair	4%	20%	16%
10% - 19%	Poor	8%	16%	12%
Negative - 9 %	Very poor	12%	8%	0%

Table 2 shows that most of the test items were able to differentiate amongst the abilities of the students. The total of items with good or better discrimination was 76% from the first grade test, 56% from the second grade test, and 72% from the third grade test.

Lastly, the total of effective distractors (good and very good distractors): there were 37 items in the first grade test, 41 items in the second grade test, and 33 items in the third grade test. Otherwise, ineffective distractors (poor and very poor distractors) have to be removed or revised, via: 47 items from the first grade test, 33 items from the second grade test, and 50 items from the third grade test. The rest, the fair distractors need to be rechecked or revised in order to become better distractors. They were: 16 items from the first grade test, 26 items from the second grade test, and 17 items from the third grade test.

In addition, 58 items were analyzed to know which items matched the intended demands of the syllabi. Based on the analysis, some mistakes were found in the construction of distractors. Probably it was these factors that made some of the distractors become ineffective. One cause included distractors which were not homogenous with the correct answer. Five such items were provided with a different phrase to the correct answer via: item number 11 in the first grade test, item number 1 in the second grade test, and items numbers 3, 18, and 24 in the third grade test. Besides, these distractors there were options which were either shorter or longer than the others such as in items numbers 7, 9, 11 and 22 in the first grade test, numbers 1, 6, 7, and 10 in the second grade test, and numbers 1, 3, 4, 9, 11, 12, 13, 17, 21, 23, and 24 in the third grade test. Next, there were items with wrong grammar as occurred in numbers 8, 13, and 21 in the first grade test, number 3 in

the second grade test, and number 24 in the third grade test. Then some key answers such as those included in items numbers 6 and 11 of the second grade summative test were provided exactly as they appeared in the texts, hence the test takers were able to answer those questions easily. Lastly, some items or options also gave a clue to the correct answer such as test items numbers 11 and 20 in the first grade test, numbers 6, 7, and 8 in the second grade test, and item number 3 in the third grade test.

Apart from all the above, all alternatives included in the summative tests for English were begun with a capital letter. Thus there was no differentiation between questions and statements in the stem. Moreover, there were many mistakes in writing either in the stem or the option. Misspelt words were found in items 13 and 19 in the first grade test, items numbers 6, 7, and 11 in the second grade test and items numbers 3 and 4 in the third grade test: the English teacher must correct these mistakes.

Discussions

Based on all the findings above related to validity, reliability, difficulty index, discrimination index, and the effectiveness of distractors, it was found that the quality of the summative tests for English at MAN Meulaboh I was on average just sufficient, except in terms of the difficulty index and the effectiveness of distractors. Yet, content analysis is the most important point that should be paid more attention because it relates to the curriculum. All teachers are required to teach based on the curriculum and to give tests based on what was taught to test the ability of their students. Furthermore, according to Fulcher and Davidson (2007) the content of the test has to be a representative sample from the domain that is to be tested. If the teachers test about material that was not appropriate for or not in the syllabus or the test does not include all the materials in the syllabus then the test is likely to be invalid.

According to Brown (2004) indirect testing means the test takers are not involved in actually performing the task itself. This is so in the first grade test items numbers 1, 2, 3, 4, 5, 6, and 7 and in the second grade test number 19. These tests were designed to measure the students' ability in speaking skills. It would be better if these test items are changed to be direct tests in which the English teacher tests the students directly with an oral test because if a teacher intends to measure speaking skills indirectly it is an inadequate way to test the

students' real speaking ability. Additionally, test items numbers 19, 20, 21, and 22 in the first grade, numbers 15, 16, 17, and 20 in the second grade, and numbers 21, 22, 23, 24, and 25 in the third grade constitute indirect tests for writing skills. They consist of grammar that is attached as a language feature in the reading text or as a direct test for writing skills. The teacher should measure students' writing ability by asking them to write by using a proper grammatical structure.

Concerning validity coefficient, an item is judged valid if the item score has significant correlation with the total score. Validity coefficient can be calculated through point bi-serial correlation or product moment. If the total of the item is 25 items of each grade, and 5% (0.05) level of significance is used, so the *t* (table) is 0.381. This means that an item is called valid when the item validity coefficient is higher than 0.381.

Concerning the degree of reliability, Linn and Gronlund (2002) say that the reliability index of a teacher made test usually varies between 0.60 and 0.85. Therefore, the tests for the first grade and for the second grade reached a high reliability index, and the test for the third grade which had a quite high reliability index were all good. This means that the summative tests for English at MAN Meulaboh I had a consistent degree of reliability.

According to Brown (2004) both very difficult and very easy items cannot differentiate between upper groups and lower groups when testing. In other words, a good test consists of neither very easy nor very difficult items. In this case, Asaad and Hailaya (2005) argue that for most teacher-made tests, only 0.30 - 0.70 on the difficulty index would be acceptable. Therefore, the items that have to be saved in the first grade are test items 3, 5, 6, 7, 8, 11, 12, 14, 17, 19, 20, 21, 22, 23, 24, and 25, in the second grade are test items 2, 7, 9, 12, 13, 15, 16, 17, 18, 19, 23, and 25, and in the third grade are test items 14 and 22.

Furthermore, the higher the score index a test item has, the higher the discrimination index of the test item is. In other words, English teachers should save test items having an excellent or a good index of discrimination. If the test item is fair in the discrimination index, it means that the item is acceptable but it needs improvement while if they are poor, the items must be revised or discarded. Brown (2004) also says that the items that scored near zero have to be discarded because they have a very poor discrimination index. Thus, there were five items which must be discarded. They include three items from the first grade test viz: numbers 1, 8 and 16 and two items from the second

grade test - numbers 14 and 15. Apart from that, nine items must be revised; they are test items numbers 4 and 9 from the first grade test; numbers 1, 2, 10, and 21 from the second grade test, and numbers 18, 21, and 25 from the third grade test. Finally, there are ten test items which are acceptable but should be improved, namely test item number 12 from the first grade test; items numbers 6, 11, 13, 15, and 18 from the second grade test and items numbers 1, 2, 3, and 7 from the third grade test.

Related to the effectiveness of distractors, according to Gronlund and Waugh (2009) if the stem is in question form, each alternative will begin with a capital. Otherwise, a lowercase letter is used to start each alternative when the stem is in an incomplete statement form.

CONCLUSIONS AND SUGGESTIONS

Conclusions

Based on the results of the analysis of the validity of the summative test for English at MAN Meulaboh I, it can be concluded that the validity of the test was not good since the percentage in content validity, construct validity, and validity coefficient was below 73%. Secondly, the summative test for English from MAN Meulaboh I had a consistent degree of reliability because all the tests had a high reliability index: 0.70 for the first and the second grade tests and 0.88 for the third grade test. Thirdly, 60% of the items in the summative test for English for the first grade, 48% of items for the second grade test, and 8% of items for the third grade were acceptable. This meant that the third grade had a very poor difficulty index and needs to be revised. Fourthly, the discrimination index for the summative test for English was good ranging between 72% and 76%. In detail, good discrimination index in the first grade test items was 76%, in the second grade was 56%, and the third grade was 72%. Hence, more than a half of the discrimination index from the summative test for English was up to standard. Finally, the quality of distractors in the third and the first grades was not good. The total of poor distractors in the third grade was 50% and in the first grade was 47% whilst in the second grade, only 33% of the distractors were poor.

Suggestion

All English teachers need to know how to analyze their test items and how to do analysis of test items in order to improve the quality of

summative tests for English. The teachers also have to learn how to conduct the tests of tests by themselves without taking the tests they use directly from books or from the internet. Besides, related to content validity, the English teachers or makers of summative tests for English must ensure all basic competencies for all syllabus materials are included in their summative tests for English to ensure that their students are assessed based on the intended demands of the syllabus to make their tests or their judgments more accurate. Apart from that, the index of difficulty of their tests must also be considered because it relates to the discrimination index. If an item is very difficult, that means that that item is not able to discriminate between upper and lower groups of students. In short, teachers should provide neither too easy nor too difficult items, so that each item will meet an ideal difficulty index. Either easy or difficult items can be affected by the effectiveness of distractors. Therefore, teachers should also pay attention to the quality of distractors. They have to be cautious in constructing distractors in order that various common mistakes are avoided.

REFERENCES

- Asaad, A., & Hailaya, W. M. (2005). *Measurement and Evaluation: Concept and Principles*. Quezon: REX Printing Company, Inc.
- Asia E-University. (2009). *Classroom Assessment: Chapter 10 in Educational Psychology (online course)*. Retrieved January 17, 2005 from http://www.peoplelearn.homestead.com/beduc/chapter_10.pdf
- Basanta, C. P. (2012). Coming to grips with progress testing: some guidelines for its design. *A Journal from English Teaching Forum*, 50(3), 37-40.
- Brown, H. D. (2004). *Language Assessment: Principles and Classroom Practices*. White Plains, N. Y.: Pearson Education.
- Burton, S. J., Sudweeks, R. R., Merrill, P. F., & Wood, B. (1991). *How to Prepare Better Multiple Choice Tests: Guidelines for University Faculty*. Retrieved from <https://testing.byu.edu/handbooks/betteritems.pdf>
- Derrick, D. (2013). Teaching beyond the test: A method for designing test preparation classes. *English Teaching Forum*, 51(4), 20-27.
- Fulcher, G., & Davidson, F. (2007). *Language Testing and Assessment*. New York: Routledge.

- Fulcher, G., & Davidson, F. (2013). *The Routledge Handbook of Language Testing*. New York: Routledge.
- Gronlund, N. E., & Waugh, C. K. (2009). *Assessment of Student Achievements*. New York: Pearson Education.
- Harlen, W. (2005). Teacher's summative practice and assessment for learning - tensions and synergies. *The Curriculum Journal*, 16(2), 207-223.
- Harrison, A. (1991). *A Language Testing Handbook*. Hong Kong: Macmillan.
- Heaton, J. B. (2000). *Writing English Language Tests*. Beijing: Foreign Language Teaching and Research Press.
- Hughes, A. (2003). *Testing for Language Teachers*. Cambridge: Cambridge University Press.
- Khalifa, H., & Weir, C. J. (2009). *Studies in Language Testing*. Cambridge: Cambridge University Press.
- Liao, Y. (2004). Issues of validity and reliability in second language performance assessment. *Columbia University Working Papers in TESOL & Applied Linguistics*, 4(2), 1-4.
- Linn, R. L., & Gronlund, N. E. (2002). *Measurement and Assessment in Teaching*. Upper Saddle River, N. J.: Prentice Hall.
- McCowan, R. J., & McCowan, S. C. (1999). *Item analysis for criterion-referenced tests*. New York: Center for Development of Human Services.
- Okunya, L. O. (2014). Validity and reliability of teacher-made tests. *African Educational Research Journal*, 2(2), 61-71.
- Popham, W. J. (2004). Curriculum, instruction, and assessment: Amiable allies or phony friends?. *Teachers College Record*, 106(3), 417-428.
- Roegier, D. (2014). Assessment literacy: Building a base for better teaching and learning. *English Teaching Forum*, 52(3), 2-13.
- Roszkowski, M. J., & Sprent, S. (2011). Issues to consider when evaluating "tests". In *Financial planning and counseling scales* (pp. 13-31). Springer New York.
- Sunarya. (2003). *Panduan Anates Kr-2003*. Retrieved July 21, 2015 from http://file.upi.edu/Direktori/FIP/JUR._PSIKOLOGI_PEND_DAN_BIMBINGAN/195911301987031-YAYA_SUNARYA/BAHAN_EVALUASI-ASESMEN/ANALISIS.pdf

- Thomas, J., Allman, C., & Beech, M. (2004). *Assessment for the Diverse Classroom: A Handbook for Teachers*. Retrieved January 20, 2015 from http://www.fldoe.org/core/fileparse.php/7690/urlt/0070083-assess_diverse.pdf
- Walker, C., & Schmidt, E. (2004). *Smart Tests: Teacher-Made Tests That Help Students*. Markham, O. N.: Pembroke.