VALIDITY AND RELIABILITY OF ENGLISH SUMMATIVE TEST ITEMS DESIGNED FOR SMP STUDENTS IN DEPOK

YON A.E

STKIP PANCA SAKTI BEKASI
yon _amrizal@yahoo.co.id

ABSTRACT

The focuss of this research is to know how well the students can master basic competences in the school-based curriculum. The test result is used as inputs to make decisions and judgments about the students' progress and remedial programs. In fact, the test was not analyzed before being given to the test takers so that there was no information about the test quality. This was an evaluation research. The English test were analyzed by using ANATEST program version 4.0.3 to get the empirical data about the characteristic of each item. The result of the analysis showed the weaknesses of the test. The test was valid in the content because all basic competences which were taught have been covered in the test items. However, the test has the weaknesses in terms of the suitability of item indicators in the test blueprints with the test items. The test reliability was in the level of criteria "low" because it was found the coefficient index was 0.25. In terms of the difficulty indices, it was found 52 % items were in the level of criteria "easy". In terms of the discrimination indices, 90 % items were in the level of criteria "poor". In terms of the function of distractors, 90,7 % distractors in the test items did not function.
Keywords: Validity, Reliability, English Summative Test Items

ABSTRAK

Tujuan dari penelitian ini adalah untuk mengetahui sejauh mana siswa dapat menguasai kompetensi dasar dalam kurikulum. Hasilnya digunakan guru sebagai masukan untuk membuat keputusan dan untuk mengetahui kemajuan siswa dalam belajar serta untuk mengetahui perlu atau tidak diadakan program remedial. Kenyataanya, tes tidak dianalisis terlebih dahulu sebelum diberikan kepada siswa sehingga tidak diketahui kualitas sebuah tes. Penelitian ini merupakan penelitian evaluasi yang dianalisis dengan menggunakan program ANATEST versi 4.0.3 untuk mendapatkan data empiris tentang karakteristik masing-masing item. Hasil analisis menunjukkan kelemahan dari tes. Tes valid dalam konten karena semua kompetensi dasar yang diajarkan telah diujikan dalam item tes. Namun, tes memiliki kelemahan dalam hal kesesuaian indikator item tes. Uji reliabilitas diemukan pada kriteria "rendah" karena ditemukan indeks koefisien adalah 0,25. Dalam hal indeks kesulitan, ditemukan 52% item yang di tingkat kriteria "mudah". Dalam hal indeks diskriminasi, 90% item yang di tingkat kriteria "jelek". Dalam hal fungsi distraktor, 90,7% distractors di item tes tidak berfungsi.
Kata kunci: Validitas, Reliabilitas, Item Tes sumatif Bahasa Inggris

INTRODUCTION

Education Law No. 20/2003 verse 58 article 1 requires teachers to evaluate their students' achievement in order to get information about students' progress. This means that evaluation toward students' achievement is one of the teachers' duties. The results of evaluation can be used as inputs to

improve the quality of teaching programs for students' benefits.

In relation to this, the local government of Depok through the educational department has conducted a summative test, a local examination held at the end of semester at schools in Depok. The examination is held at schools by using a multiple-choice test. This test is designed by some selected teachers from every subject. All items are made on the basis of indicators in the blue prints of the test.

However, when the test items were tested to the students in the local examination of 2012/2013 academic year, the writer found some problems on the test items. Based on the writer's observation during the test, it was found that pictures displayed on the test papers were not clear, some options used in the test items were ambiguous because they seemed to have more than one correct answer.

Ideally, a language test which will be used as an instrument to measure students' achievement should be good. To ensure a language test is good or not, an item analysis can be applied. Item analysis involves the calculation of difficulty indices, discrimination indices, and function of distractors of the test items in the case of multiple-choice test as stated by Hughes (2002, p. 225). The results of the analysis can be used by test designers as inputs to maintain the good items and to revise the bad items for better test construction..

The problems appeared because no analysis was applied to the test before being given to the test takers so that there was no information about the quality of the English test. Analysis on the test is important to do to ensure whether the test has fulfilled the criteria of a good test. This can be done by analyzing the difficulty indices, item discriminations, function of distractors, reliability and validity of the English test.

Based on background of the problem above, the focus of the research is the Validity and Reliability Analyses on the Language Testing Items of SMPN of Grade VIII in Depok in academic year 2012/ 2013. The researcher was focused to know the difficulty indices, item discriminations, function of distracters, reliability and validity of the English test. These aspects were judged by using criteria suggested by experts.

LITERATURE REVIEW

Some experts have proposed the definitions of test. Bachman (1990, p. 20) said that test is a measurement instrument designed to elicit specific sample of an individual's behavior. Then Brown (2004, p. 3) defines test as a method of measuring a person's ability, knowledge, or performance in a given domain. Hughes (2002, p. 4) confirmed that test is needed in order to provide information about achievement of groups of learners, without which it is difficult to see how rational educational decisions can be made.

A good test designer should consider some characteristics of a good test to make the test effective. In relation to this, Weir (1993, pp. 22-23) explained that there are some characteristics of a good test. They are as follows:

a) A good test should test what the test designer wants it to test. In other words, a good test should be valid.
b) A test should be consistent in the results. In other words, a good test should be reliable.
c) A test should be efficient in cost and easy to construct.

d) A good test should be able to discriminate between the performances of candidates at different levels at attainment.
e) A good test should include the full range of appropriate skills and abilities, as defined by the objectives of syllabus and course book unit.
f) The test tasks should be unambiguous, giving a clear indication of what examiner is asking.
g) A reasonable amount of time must be provided for the majority of the test takers to be able to complete the test tasks.
h) A good test should have a clear format and layout of the questions.

Brown and Hudson (2002, p. 113) defined item analysis as the systematic statistical evaluation of the effectiveness of individual test items. In this case, item analysis is usually done for purposes of selecting which items will remain on future revised and improved versions of the test. Then Arikunto (2009, p. 205) defined item analysis as a systematical process which will give particular information about the strengths and weaknesses of the items made by test designers. Hughes (2002, p. 225) stated that the purpose of item analysis is to examine the contribution that each item is making.

In order to analize the test the researcher used the formulas suggested by the expert, they are:

1. Difficulty Indices

Finocchiaro and Sydney (1983, p. 297) defined difficulty indices as a judgment regarding the ease or difficulty of test items through an analysis of the number of students who answered them correctly. To estimate the difficulty indices of items, Henning (1987, p. 49) suggested the formula below:

$$\boldsymbol{P = \frac{\Sigma Cr}{N}}$$

Note:
P = difficulty indices
ΣCr = the sum of correct responses
N = the number of examinees

To interpret the index of difficulty indices, Hughes (2002, p. 225) clarified that items with the difficulty indices index of 0.00 - 0.30 are classified into difficult items, the index of 0.30 - 0.70 are classified into moderate items, and the index of 0.70 - 1.00 are categorized as easy items.

2. Discrimination Indices

Hughes (2002, p. 226) noted that a discrimination index is an indicator of how well an item discriminates between weak candidates and strong candidates. In this case, the higher its discrimination index, the better the item discriminates between high students and low students.

To estimate the discrimination indices of items, Arikunto (2009, p. 213) suggested the formula below:

$$\mathbf{D = \frac{Ba}{Ja} - \frac{Bb}{Jb}}$$

Note:
D = discrimination indices
Ba = the number of correct responses in the high group
Bb = the number of correct responses in the low group
Ja = the number of test takers in the high group
Jb = the number of test takers in the low group

Arikunto (2009, p. 211) defines discrimination indices as the items' ability to discriminate between high

students and low students. Moreover, Arikunto (2009, p. 218) classified the index of discrimination (D) into the criteria below:

a) D = 0.00 – 0.20 (Poor items)
b) D = 0.20 – 0.40 (Satisfactory items)
c) D = 0.40 – 0.70 (Good items)
d) D = 0.70 – 1.00 (Excellent items)
e) D = < 0.00 (Very poor items)

3. Distractors

Distractors are incorrect options in multiple-choice items. They influence the quality of items. Hughes (2002, p. 228) described that it is necessary to analyze the performance of distractors since the distractors that do not work make no contribution to test reliability. To determine whether a distractor functions well or not, Arikunto (2009, p. 220) clarifies that disctractors can be considered good if they are chosen by at least 5% of all test takers. In other words, the functional disctractors should be chosen by at least 5% of test takers.

4. The Reliability of Test

A good test should be reliable. Weir (1990, pp. 31-32) stated that reliability refers to whether the results of a test can be produced consistently. It means a reliable test must be consistent in the results. Then Gay and Airasian (2003, pp. 141-145) divided the reliability of test into five types of reliability: stability, equivalence, equivalence and stability, internal consistency, and rater agreement. Stability (test-retest reliability) is the degree to which scores on the same test are consistent over time. Equivalence is alternative forms of a test which should be identical in variable, structure, difficulty level, scoring, etc. Equivalence and stability is the form of reliability which combines equivalence and stability. Internal consistency is the form of reliability that deals with the consistency among the items of one test at one time.

To measure the reliability of a test more easily, the formula of Kuder-Richardson 21 (KR-21) can be used. For this, Gay and Airasian (2003, p. 144) stated that KR-21 is easier to use since it requires less time than any other method of estimating reliability. The formula is below:

$$r_{total\ test} = \frac{(\mathrm{K})(SD^2) - \bar{x}\ (K - \bar{x})}{(SD^2)(K-1)}$$

Note:

K = the number of items in the test
SD = the standard deviation of the scores
$\bar{X}$ = the mean of the scores

The extent of reliability of a test is determined by the index of reliability coefficient. In relation to this, Ary, Jacobs, and Asghar (2002, p. 262) described that achievement tests with the reliability coefficient of ≥ 0.90 are judged to have high reliability. Moreover, the index ≥ 0.60 - 0.70 are regarded as moderate reliability. Based on the theory, it can be formulated that the reliability coefficient for achievement tests has the criteria below:

a) 0.00 – < 0.60 (Low)
b) ≥ 0.60 – < 0.90 (Moderate)
c) ≥ 0.90 – 1 (High)

5. The Validity of Test

A good test should be valid. Related to this, Popham (1981, p. 99) pointed out that the validity of test indicates whether the test measures

what it's supposed to measure. Weir (1993, p. 19) said that a valid test should test what the test makers want it to test. Hughes (2002, p. 26) defines validity as the criteria to measure a test whether it can measure accurately what it is intended to measure. Hatch and Lazaraton (1991, p. 540) stated that the validity of test consists of content validity, predictive validity, and face validity. Content validity has to do with how well a test can test what it purports to test. Predictive validity refers to the use of tests as valid for the predictive purposes. Then Farhady (1986, p. 24) said that face validity refers to the extent to which the physical appearance of the test corresponds to what it is claimed to measure.

RESEARCH METHODOLOGY

This chapter presents the discussion about methodology, subject and sample of the research, instrument of the research, technique of collecting the data and technique of analyzing the data

1. Purpose of the Research

Based on the research questions, the purpose of the research is as follows:

(1) To know the difficulty indices, item discriminations, function of distractors, reliability and validity of the English test for SMP students of Grade VIII in the local examination of 2012/2013 academic year in Depok

(2) To know the realibility of the English test for SMP students of Grade VIII in the local examination of 2012/2013 academic year in Depok.

(3) To know the validity of English test for SMP students of Grade VIII in the local examination of 2012/2013 academic year in Depok.

2. Type of the Research

Based on the research questions and the purpose of this research, type of the research is evaluation research. Gay and Airasian (2003, p. 7) mentioned that evaluation research is concerned with making decisions and judgment about the quality, effectiveness, merit, or value of educational programs, products, or practices. Then Mertler and Charles (2005, p. 260) stated that evaluation research is done to determine the merits of various products e.g. textbooks, tests, and instructional programs used in education. Weiss (1972, p. 1) explained that evaluation research purposes to compare "what is" with "what should be".

3. Subject and Sample

1. Subject of the Research

The Subject of the research was the English test done by SMP students of Grade VIII of semester I of 2012/2013 academic year in Depok. There were 410 SMP students who took the test.

2. Sample

The writer used the students' answer sheets as the sample of research in order to get data. To make the writer easy in collecting and analyzing data, the writer took 10% sample (10% of 410 students' answer sheets). Gay and Arasian (2003: 112) explained that in general the sample size of a research ranges from 10 to 30% sample.

4. Instrumentation

The data used in this research consisted of qualitative data and quantitative data. The qualitative data was the items used in the English test for SMP students of Grade VIII of semester I of 2012/2013. The quantitative data was the students' answers on the test and their scores. The instrument which was used in this research was the achievement test designed by SMP English teachers in Depok.

5. Technique of Collecting the Data

Mertler and Charles (2005, p. 271) stated that data in evaluation research can be gathered by collecting and analyzing the relevant documents of the unit concerns under study. Hence, in collecting the data the writer implemented the following steps:

a. The writer asked officially the necessary documents.
b. The students' answers were computed and processed by using ANATEST program version 4.0.3 to get some statistical information.
c. The writer printed the results of data computation.
d. The writer took a set of items used in the English test.

6. Technique of Analyzing the Data

1. Analyzing the Difficulty Indices

To analyze the difficulty indices of the items, the writer used the formula suggested by Henning (1987, p. 49), as follows:

$$P = \frac{\Sigma Cr}{N}$$

Note:
P = difficulty indices
ΣCr = the sum of correct responses
N = the number of examinees

The result of calculation was judged by using the index of difficulty indices proposed by Hughes (2002, p. 225) below:

a. The index of 0.00 - 0.30 are classified into difficult items.
b. The index of 0.30 - 0.70 are classified into moderate items.
c. The index of 0.70 - 1.00 are categorized as easy items.

2. Analyzing the Discrimination Indices

To analyze the discrimination indices of the items, the writer used the formula suggested by Arikunto (2009, p. 213), as follows:

$$\mathbf{D} = \frac{\mathbf{Ba}}{\mathbf{Ja}} - \frac{\mathbf{Bb}}{\mathbf{Jb}}$$

Note:
D = discrimination indices
Ba = the number of correct responses in the high group
Bb = the number of correct responses in the low group
Ja = the number of test takers in the high group
Jb = the number of test takers in the low group

The result of calculation was judged by using the index of discrimination indices proposed by Arikunto (2009, p. 218) below:

a) D = 0.00 – 0.20 (Poor items)
b) D = 0.20 – 0.40 (Satisfactory items)
c) D = 0.40 – 0.70 (Good items)
d) D = 0.70 – 1.00 (Excellent items)
e) D = < 0.00 (Very poor items)

3. Analyzing the Function of Distractors

To determine whether a distractor functions well or not, Arikunto (2009, p. 220) clarifies that disctractors can be considered good if they are chosen by at least 5% of all test takers.

4. Analyzing the Reliability

To analyze the reliability of the test, the writer used the formula of Kuder-Richardson 21 (KR-21. The formula is below:

$$r_{total\ test} = \frac{(\mathrm{K})(SD^2) - \bar{x}\ (K - \bar{x})}{(SD^2)(K-1)}$$

Note:

K = the number of items in the test
SD = the standard deviation of the scores
$\bar{X}$ = the mean of the scores

The result of calculation was judged by using the criteria of reliability coefficient for achievement tests suggested by Ary, Jacobs, and Asghar (2002, p. 262) below:

a) 0.00 – < 0.60 (Low)
b) ≥ 0.60 – < 0.90 (Moderate)
c) ≥ 0.90 – 1 (High)

5. Analyzing the validity

The test validity was analyzed in three ways. The first was by analyzing the content validity by examining whether all basic competences of English for SMP students of Grade VIII in the school based curriculum have been included in the test items. The second was by examining the suitability of item indicators in the test blueprints with the test items. The third was by examining aspects of the test appearance.

FINDINGS AND INTERPRETATION

A. Finding

To get the general description about the students' score distribution, the writer used the data of students' raw scores to be analyzed. The Data of students' score distribution can be seen in Table 1 below.

Table 1.
The Data of Score Distribution of the English Test For SMP of 2012/2013 in Depok

No.	Score (X)	Frequency (F)	Proportion (P)
1.	25	1	2.44%
2.	26	3	7. 32%
3.	27	2	4. 88%
4.	28	2	4. 88%
5.	29	5	12.19%
6.	30	1	2.44%
7.	32	5	12.19%
8.	33	5	12.19%
9.	34	2	4. 88%
10.	35	5	12.19%
11.	36	6	14.64%
12.	38	2	4. 88%
13.	39	2	4. 88%
	Total	41	100%

Mean : 32.34
Median : 33
Mode : 36
Standard Deviation : 3.90
Number of Subject : 41

The score distribution of the English test was not in normal distribution since it was found the shape of curve was negatively skewed. It was because the mean (32.34) was lower than the median (33) and mode (36). Related to this, Gay and Airasian (2003, p. 354) said that for a negatively skewed distribution the mean is always lower or smaller than the median and the median is always lower or smaller than the mode. Moreover when a score distribution is negatively skewed, it indicates that most of the test items are easy to answer. The figure also shows that the proportion of score above the mean was higher than the score below the mean because 53.66% scores were above the mean and 46.34% scores were below the mean. It means that the test takers who got score above the mean were more than those who got score below the mean. Based on the evidence, it can be said that the test takers who could answer the test items correctly were more than those who could not.

1. The Difficulty Indices

The findings of difficulty indices of the English test are based on the data analysis using ANATEST Program version 4.0.3.

The Researcher found that there were 26 easy items (52 %) in the test. There were 19 moderate items (38%) in the test. The difficult items were 5 items (10 %). From the data, it was found that the difficulty indices of the English test for SMP was dominated by easy items. 52 % of the items could be answered easily by the test takers. In other words, most of the items could be answered by the test takers easily.

2. The Discrimination Indices

The findings of discrimination indices of the English test are based on the data analysis using ANATEST Program version 4.0.3. There were 45 poor items (90 %) in the test. There were only 5 good items (10 %) in the test. From the data, it was found that the discrimination indices of the English test for SMP was dominated by poor items. It means that most of the items could not discriminate between good and poor students.

3. Function of Distractors

The distractors which function were the distractors which were chosen by at least 12 test takers (5 % of 41 subjects). The total of distractors which were analyzed was 150 distractors. In this case, each item (50 items) consisted of 3 distractors. It was found that from 150 distractors there were only 14 distractors (9.3 %) which functioned well. There were 136 distractors (90.7 %) which did not function in the test. From the data, it was found that the distractors in the English test for SMP were dominated by non functional distractors. It means that most of the distractors could not make the test takers doubt to choose the correct answers so that most items could be answered easily by most test takers.

4. Reliability

The reliability of the English test was as follows:

$$r_{total\ test} = \frac{(\mathrm{K})(SD^2) - \bar{x}\ (K - \bar{x})}{(SD^2)(K-1)}$$

$$r_{total\ test} = \frac{(50)(3.90^2) - 32.34\ (50 - 32.34)}{(3.90^2)(50-1)}$$

$$r_{total\ test} = \frac{760.5 - 571.1244}{745.29}$$

$$r_{total\ test} = 0.25$$

Based on the calculation, it was found that the reliability coefficient of the English test was 0.25. From the result, it was found that the English test has low reliability.

5. Validity

Based on the analysis, the distribution of basic competences in the test items, there were 24 items (48%) which represented Basic Competence No. 1. Basic Competence No. 2 was covered in 10 items (20%). Basic Competence No. 3 was covered in 16 items (32%). It can be said that the distribution of basic competences in the test items were dominated by Basic Competence No. l.

The result of analysis related to the suitability of item indicators and the test items can be seen in the test blueprints, which the test designers wanted to test in the test items. It was found that there were 16 item indicators which were not suitable with the test items. The item indicators which were not suitable with the test items included Item Indicator No. 7, 8, 9, 11, 14, 15, 22, 23, 24, 25, 26, 27, 28, 29, and 31.

Based on the result of analysis on the test appearance, it was found some weaknesses concerning with some aspects, as follows:

1) Clarity of Pictures
2) Inconsistency in using punctuation
3) Typing errors in numbering items

B. Interpretation

After knowing the results of analysis of the English test for SMP students of Grade VII of 2012/2013 in Depok, it was found some weaknesses of the English test.

1. Difficulty Indices

Based on the result of analysis of difficulty indices, it can be said that most test items were very easy to answer. This doesn't conform to the criteria which should be. Ideally, items for an achievement test should be made in the level of criteria "moderate" with the difficulty index of 0.30 – 0.70 as suggested by Hughes (2002, p. 225). From the result of analysis, it can be said that the level of item difficulty of the test was in the level of criteria "easy" since the majority of the test items were easy to answer correctly.

2. Discrimination Indices

The weaknesses of the test were also found in terms of the discrimination indices of the test items. Based on the result of analysis, it was found that most items (90%) could not discriminate between high and low students because most items could be answered by both high and low students. From the evidence, it can be said that discrimination indices of the English test were in the level of criteria "poor". This does not conform to what should be. Weir (1993, p. 22) said that a good test should be able to discriminate between the performances of candidates at different levels.

3. Function of Distractors

Related to the function of distractors, the result of analysis showed that most distractors (90.7 %) in the English test did not function. The distractors could not make the test takers doubt to choose the correct answers. If distractors do not function, the test items will be easy to answer. This will make the item discrimination poor. Related to this, Hughes (2002, p. 228) said that distractors which do not function make no contribution to test reliability.

4. Reliability

In terms of the test reliability, it was found that the reliability coefficient

of the English test was 0.25. It can be said that the test reliability was in the level of criteria "low". It means that the English test has low reliability as the instrument to measure the students' achievement. A test with low reliability can be said that the test tends to be inconsistent in the results. Moreover, this is not suitable with what should be. For achievement tests, test reliability should be in the level of criteria "high" with the index of ≥ 0.90 – 1 as suggested by Ary, Jacobs, and Asghar (2002, p. 262).

5. Validity

In the case of test validity, the English test can be judged "valid" in the test content because the empirical data showed that all basic competences in the school based curriculum which were taught to the students were covered in the test items. However, the result of analysis on the suitability of item indicators with the test items showed the weaknesses. It was found that some of the item indicators made by the test designers in the test blueprints were not suitable with the test items. The weaknesses were also found in terms of the test appearance including clarity of pictures, inconsistency in using punctuation, and typing errors in numbering items.

CONCLUSION AND SUGGESTION

A. Conclusion

After analyzing the English test and getting the results, the writer concludes that:

1. The difficulty indices of the English test are in the level of criteria "easy" as it was found 52 % of the test items could be answered easily. In terms of the discrimination indices, the discrimination indices of the English test are in the level of criteria "poor" as it was found 90 % of the test items could not discriminate between the high and low students. Concerning the function of distractors, most distractors do not function as it was found 90.7 % of the distractors in the test items do not function.
2. The test reliability is low as it was found the index of reliability coefficient is 0.25. In terms of the test validity, the test can be judged "valid" in the content as it was found all basic competences which were taught have been covered in the test items.

As a whole, it can be concluded that the English test is not good yet. It still needs improvements. Hence, the English test should be revised.

B. Suggestion

Based on the research findings, the writer wants to provide some suggestions, as follows:

1. English teachers should pre-test, analyze, and revise the tests which they design before being tested to their students.
2. Education Department (Diknas) and schools in Depok need to invite the practitioners of evaluation or language testing to train the English teachers about how to design a good test.
3. English Teachers Forum or "Musyawarah Guru Mata Pelajaran" (MGMP) in Depok should be more active in preparing and discussing tests.
4. It is expected that other researchers will conduct further research on the evaluation of test and test design.

REFERENCES

Arikunto, S. 2009. *Dasar-Dasar Evaluasi Pendidikan* (Edisi Revisi). Jakarta: Bumi Aksara.

Bachman, L F. 1990. *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Best, J and J. Khan. 2003. *Research in Education* (9th ed.). Boston: Pearson Education Company.

Brown, H. D. 2004. *Language Assessment: Principle and Classroom Practice.* San Fransisco: Longman.

Brown, J. D and T. Hudson. 2002. *Criterion-Referenced Language Testing.* Cambridge: Cambridge University Press.

Burton, S. J, R. R. Sudweeks, P.F. Merrill and B. Wood. 1991. *How to Prepare Better Multiple-Choice Test Items: Guidelines for University Faculty*. Brigham: Brigham Young University Testing Services.

Finocchiaro, M and S. Sako. 1983. *Foreign Language Testing: A Practical Approach*. New York: Regents Publishing Company, Inc.

Gay, L.R and P. Airasian. 2003. *Educational Research: Competencies for Analysis and Application* (7th ed.). New Jersey: Pearson Education, Inc.

Henning, G. 1987. *A Guide to Language Testing: Development-Evaluation-Research.* Massachusetts: Newbury House Publishers.

Hughes, A. 2002. *Testing for Language Teachers* (2nd ed.). Cambridge: Cambridge University Press.

Mertler, C. A and C. M. Charles. 2005. *Introduction to Educational Research* (5th ed.). New York: Pearson Education, Inc.

Richards, J. C. 2001. *Curriculum Development in Language Teaching.* Cambridge: Cambridge University Press.

Richard, P and L. Amato. 2003. *Making It Happen: From Interactive to Participatory Language Teaching-Theory and Practice* (3rd ed.). New York: Pearson Education, Inc.

Weir, C. J. 1990. *Communicative Language Testing*. London: Prentice Hall International (UK) Ltd.

Weir, C. J. 1993. *Understanding and Developing Language Tests*. London: Prentice Hall International (UK) Ltd.

---------- 2010. *Juknis Analisis Butir Soal*. Jakarta: Direktorat Pembinaan SMA.