Teacher-made Summative Test: An Analysis of Test Format, Index of Difficulty, Discrimination, and Distractors

Isabell Sengkaton[1)]
Muslih Hambali[2)]
Soni Mirizon[3)]
smirizon@gmail.com

Abstract: This study was aimed at finding out the quality, index of difficulty, discriminating power, and effectiveness of distractors of teacher-made English summative test item format. The sample of this study was 44 eleventh-grade students of one public senior high school in Palembang. The data were collected from the result of the English summative test responded by eleventh-grade students of the school. The data were analyzed by using the table of multiple-choice test item format and item analysis. The result of the test item format showed that none of the item was in the good (accepted) category, 34 (85%) items were in the moderate category, and 6 (15%) items were in poor category. The result of index difficulty showed that 10 (25%) items were in difficult category, 12 (30%) items were in easy category, and 18 (45%) items were in satisfactory category. The result of discriminating power showed that 16 (40%) items had poor discriminating power, 10 (25%) items had moderate discriminating power, and 14 (35%) items had good discriminating power. The result of the effectiveness of distractors showed that from a total of 160 distractors 41% of them functioned effectively, 51% of the distractors functioned less effectively, and 8% of the distractors were ineffective. In short, the teacher-made English summative test of the eleventh-grade students of the school was not acceptable to be used as an instrument to evaluate the students' English learning due to many aspects of a good test in terms of quality, index of difficulty, discrimination power and effectiveness of distractors were not fulfilled.

Keywords: *test item analysis, teacher-made summative test, English*

Abstrak: Studi ini bertujuan untuk mengetahui kualitas, indeks kesulitan, kekuatan pembeda, dan keefektipan pengecoh terhadap format butir tes sumatif Bahasa Inggris. Jumlah sampel sebanyak 44 siswa kelas 11 Sekolah Menengah Atas Negeri 3 Palembang. Data studi ini merupakan jawaban hasil sumatif tes yang dikerjakan oleh ke 44 siswa tersebut dan dianalisa menggunakan tabel format butir tes berbentuk pilihan ganda. Hasil format butir tes menunjukkan tidak satupun butir soal berkatagori bagus. Dari sejumlah 40 soal, 34 (85%) butir soal tergolong sedang dan sisanya 6 soal (15%) termasuk katagori kurang. Dari hasil indeks kesulitan butir soal terdapat 10 butir (25%) dianggap sulit, 12 butir (30%) mudah , dan 18 butir (45%) tergolong memuaskan. Untuk kekuatan pembeda, terdapat 16 butir (40%) soal kurang, 10 butir (25%) butir sedang, dan 14 butir (35%) dianggap bagus. Sementara itu, hasil keefektipan pengecoh menunjukkan 41% dari total 160 pengecoh efektif, 51% kurang efektif, dan 8% tidak efektif. Secara ringkas tes sumatif Bahasa Inggris buatan guru untuk kelas 11 kurang dapat diterima sebagai instrumen untuk mengevaluasi pembelajaran bahasa Inggris siswa dikarenakan terdapat beberapa aspek sebagai persyaratan tes yang bagus tidak sempurna terpenuhi.

Kata-kata kunci: *analisis butir tes, tes sumatif buatan guru, Bahasa Inggris*

[1) 2) 3)] *Lecturers of Sriwijaya University*, *South Sumatera*

Evaluation has an important role in teaching and learning activity. Evaluation and the teaching and learning process are interrelated and cannot be separated. According to Djiwandono (2011), evaluation is a standard process to accumulate information regarding the teaching and learning process. Even though evaluation focuses only on the students, teachers also participate in evaluation activity. Mardapi (2008) points out that evaluation is an activity to increase the quality, performance, and productivity of an organization. In other words, by doing evaluation the teachers will have a parameter to measure whether the teaching and learning process is successful as it has been planned or not. One of the evaluation instruments commonly used is a test.

There are many types of tests to evaluate students. Djiwandono (2011) defines four types of tests based on the educational implementation of evaluation e.i formative test, summative test, pretest, and posttest. Summative test roles as a benchmark of the students' achievement after a long time treatment by the teacher from a specific subject, in this case by the English teacher. At school, teaching and learning activities usually use formative and summative tests. According to Djiwandono (2011) summative test is a test that is given at the end of the course or semester. Norman (1965) states that a summative test is designed to determine the level to which the instructional goals have been achieved, and the test also can be used to specify course grades for asserting students' acquisition of intended learning outcomes. From the explanation above it can be said that summative test is given occasionally to ensure the students' comprehension of the materials.

The learning materials that are tested in a summative test must be based on and in line with the syllabus and curriculum. It is due to the success in achieving the instructional objectives is stated in the syllabus and curriculum. As a tool to measure students` achievement, a test has some required aspects to have a corresponding result. In fact, sometimes the content of the test is not appropriate with what is stated in the syllabus, there are some mistakes in the test commonly made by the teacher. In Indonesia context, a study by Husna and Fachrurrazy (2012), reported that there were still many English teachers at elementary schools ignorant on how to design a good test. Another study conducted by Shomami (2013) proved that the distractors of the English summative test for second-grade students were poorly made where almost 83% of the items were considered ineffective as a distractors. In addition, Kristiana (2014) proved that the content validity level of the English summative test for the second-grade students was poor where almost 46.7% of the indicators were not represented in the test item. In international context, Simsek (2016) analyzed 6,450 test items made by 120 instructors (62 teachers and 58 trainers) in various fields of learning and it was found that those school teachers and trainers made similar mistakes in test construction. Another study conducted by Kurebwa and Nyaruwata (2013) at Gweru Urban Schools in Zimbabwe showed that teachers in Gweru Urban Schools could not design and construct good tests. They believed the problems occurred because of teachers' lack of competence in test construction. Anas (2019) also reported that at the University of Sunderland the average summative test score of the student was extremely low due to teachers inability to construct good summative test. These studies indicated that many teachers failed in constructing good test to measure student learning. In relation to this, Henning (2012) argued that there are four common mistakes made by the teacher in test construction: general examination characteristics, item characteristics, test validity concerns, and administrative and scoring issues.

Summative test is frequently given to measure the students' achievement after a long-time treatment by the teachers so that they can specify the score of the test taken by the students. The test is held at the end of the semester and is usually made by the teachers. The test is usually made in the form of multiple-choice and essay tests. In test provision, teachers of the specific major of the study are working together in constructing the test to fulfill the appropriateness of the test so that the test is suitable for each level of the class.

To identify whether the test has met the standards of a good test, the teacher should ideally analyze the quality of the test item. Item analysis is rewarding for teachers to improve their skills in test construction and recognizing specific areas of course content that need greater emphasis or conspicuousness. According to Downie and Health (1974), the characteristics that

determine an item analysis test are item difficulty, item discriminator, and item distractor. The item difficulty means the level of difficulty for each item test for students. The item discriminator tells how well each item test differentiates the comprehension ability among the higher and the lower students. Lattermost, item distractor indicates how effective each alternative or option for an item on multiple-choice questions.

The main objective of this study was to find out whether teacher-made English test items were well-constructed and appropriate to the students' level. In specific, this study was aimed at finding out: (1) the quality of teacher-made English summative test item format, (2) the index of difficulty of teacher-made English summative test item format, (3) the discriminating power of teacher-made English summative test item format, and (4) the effectiveness of distractors of teacher-made English summative test item format used in the English test in one public senior high school in Palembang.

METHOD

This was descriptive quantitative study. It was conducted in a public senior high school in Palembang. Forty-four eleventh grade students (25% of 177 students) of the school were chosen as the sample of the study.

The instrument used to collect the data was a teacher-made English summative test in multiple choice questions. The data of the study were taken from the results of teacher-made English summative test that those students took.

In analyzing the data, two raters were chosen to score the quality of this teacher-made English summative test items. First, the test items were analyzed using Ghofur's (2004) table of multiple choice test item formats in terms of their quality and classified into good (accepted), moderate (revised), and poor (refused) based on the material, construction and language aspects of each items. Then, the data were calculated using item analysis formula to determine each level of the items in terms of index difficulty, discriminating power and distractor effectiveness. Finally, the test items category, index of difficulty, discriminating power, and distractor effectiveness were determined.

FINDINGS AND INTERPRETATION

The Result of Item Format

An item is classified as a good if it complies of materials, constructions, and language aspects. It is said moderate if it complies one or a couple of materials, construction and language aspects and it is said poor if it does not satisfy the three criteria of material, construction and language aspects. Based on the data analysis of 40 multiple choice items, it was found that 85% items were in moderate category, 6% were in poor category, and no item was in good category, as stated in Table 1.

Table 1. The Quality of Test Item Format

Item Number	Criteria	Total Number	Percentage
None	(Accepted) Good	0	0%
1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 16, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 31, 32, 34, 35, 36, 37, 38, 39, 40	(Revised) Moderate	34	85%
3, 15, 17, 29, 30, 33	(Refused) Poor	6	15%

Out of forty questions, 85% of the items needed to be revised, while the other 15% had to be refused because those items did not fulfill all the material, construction, and language aspects.

The Result of Item Analysis

The item analysis was focused on 40 items. It was meant to find the index of difficult, discriminating power and the effectiveness of the distractors of each item.

Index of Difficulty

The level of index difficulty is about 0.00 until 1.00. An item with 0.00 difficulty level shows that it is very difficult and the item with 1.00 difficulty level shows that it is very easy. The result of index difficulty of teacher-made summative English test for eleventh-grade students is shown in table 2.

Based on the index of difficulty of the teacher-made English summative test, it can be seen that 10 (25%) items were considered

difficult, 18 (45%) items were considered satisfactory, and 12 (30%) items were easy.

Table 2. The Index of Difficulty of Test Item

Index Difficulty	Item Number	Total	Percentage
Difficult (0.00 – 0.30)	5,8,14,15,16,19,26, 34,36,38	10	25%
Satisfactory (0.30 – 0.70)	1,3,4,6,7,9.10,11,12,17,20,21,23,28,29,30, 31,33	18	45%
Easy (0.7 – 1.00)	2,13,18,22,24,25,27, 32,35,37,39,40	12	30%

Discriminating Power

The discriminating power is categorized by four levels—poor, satisfactory, good, and excellent. The result of discriminating power of teacher-made summative English test for eleventh-grade students is shown in table 3.

Table 3. Discriminatiang Power of Test Item

Discriminating Power	Item Number	Total	Percentage
Poor (0.00 – 0.20)	2,5,8,14,15,16,19,21,26,27,32,34,35, 36,38,40.	16	40%
Satisfactory (020 – 0.40)	1,3,13,17,18,22,25,33,37,39.	10	25%
Good (0.40 – 0.70)	4,6,7,9,10,11,12,20,23,24,28,29,30, 31.	14	35%
Excellent (0.70 – 1.00)	--	--	0%

The analysis of discriminating power revealed that 16 (40%) items were in poor category, 10 (25%) items were in satisfactory category, 14 (35%) items were in good category, while none of the item was in excellent category of discriminating power.

The Effectiveness of Distractors

The teacher-made English summative test consisted of 40 multiple-choice items with five possible answers. Each item had one correct answer and four distractors. In total, there were 160 distractors of the test. According to Arikunto (2013), distractor can be considered effective if it attracts more than 5% of total test takers who tried to answer the items. The distractor should be chosen more by the test takers in the lower group than the upper group. The distractor is considered ineffective if it attracts more students in the upper group. The result of the effectiveness of the distractors can be seen below.

Table 4. The Effectiveness of the Item Distractors

No	Answer	Effectiveness of distractors			No	Answer	Effectiveness of distractors		
		Effective	Less Effective	Ineffective			Effective	Less Effective	Ineffective
1	D	B	A,B,C,E	-	21	D	A,C	B	E
2	A	-	B,C,D,E	-	22	B	A,E	C	D
3	C	A,B	D,E	-	23	E	A,C	B,D	-
4	B	C,D	A,E	-	24	B	D,E	A,C	-
5	E	A,B,,D	-	C	25	A	B,D	C,E	-
6	E	A,B	C,D	-	26	A	B	C,D	E
7	E	A,B,C	D	-	27	B	E	A,C,D	-
8	B	A	C,D,E	-	28	E	A,C,D	B	-
9	B	A,D	C,E	-	29	C	A,B,D	E	-
10	D	A,C	B,E	-	30	A	B,D,E	-	C

11	D	A,B	C,E	-	31	D	A,B	C,E	-
12	D	C,D	A,E	-	32	A	-	B,C,D,E	-
13	A	C	B,D,E	-	33	E	-	B,C,D	A
14	C	E	A,D	B	34	D	A	B,E	C
15	B	C,D	A	E	35	A	-	B,C,D,E	-
16	D	A,B,C,E	-	-	36	C	-	A,B,E	D
17	A	C,D	E	B	37	A	B,C	D,E	-
18	C	A,B,D	E	-	38	B	C	D,E	A
19	C	A,D	B	E	39	B	E	A,C,D	-
20	B	D	A,C,E	-	40	B	C	A,D,E	-
					TOTAL (%)		**66 (41%)**	**81 (51%)**	**13 (8%)**

Based on the data in Table 4, it is apparent that 66 (41%) distractors were effective, 81 (51%) distractors were less effective, and 13 (8%) distractors were ineffective because it attracted more students from the upper group to choose rather than the ones in the lower group.

Interpretation of the Findings

As mentioned above, the objectives of this study were to find out the quality, index of difficulty, discriminating power, and effectiveness of distractors of the teacher-made English summative test of the eleventh-grade students of one public senior high school in Palembang. Based on the findings from the analysis of the test, some interpetations are made.

First, In relation to the quality of the test, Ahmann and Glock (1967) pointed out that item analysis is double-checking each test item to discover its quality. BSNP (2010) stated that a good test item complies of material, construction and language aspects. If an itemsonly complies one or a couple of the criteria—materials, construction, and language aspects, the item is said to have medium quality and the item should be revised. If the item does not satisfy the majority of the three criteria, the item is considered as a poor item and should not be used as an evaluation instrument. The analysis result showed that no item was in the good (accepted) category, 34 items were in medium (revised) category, and 6 items were in poor (refused) category. It means that the items in the teacher-made English summative test cannot be used as an evaluation tool for students until it was revised.

Second, referring to the index of difficultysof the test, Arikunto (2013) argued that a question is considered good when the level of difficulty of the item is moderate. The result of the analysis showed that from 40 multiple-choice items, 10 (25%) items were categorized as difficult and 12 (30%) items were catagorized as easy, while 18 (45%) items were categorized as satisfactory. An item is in the satisfactory category if the index of difficulty is not too easy nor too difficult. The findings of this study were in line with the what Thompson and Levitov (1985) stated that to calculate the ideal index of difficulty is to identify the point on the difficulty scale midway between easy and difficult items. So, 22 out of 40 items needed to be revised to meet the satisfactory index of difficulty in order that those items can be used as an evaluation instrument.

Third, concerning with the discriminating power of a test, Sudijono (2011) pointed out that discriminating power is the aptitude of an item of achievement test to be able to distinguish between the students with a high and low capability.eThe higher the result of discriminating powersof a test item,sthe moresability of testsitem to distinguishestudents whoamaster the material with students who do not master the material.sThe result of the analysis showed that 16 (40%) items were considered poor in discriminating power, 10 (25%) items were in satisfactory level, and 14 (35%) items categorized as good in discriminating power. As what Ebel and Frisbie (1986) argued that good items have a discriminating power of 0.40 and higher and poor items less than 0.20. Although 60% of the items had satisfactory and good discriminating power, the other 40% were still needed to be revised.

The last, in regard to the effectiveness of distractors of a test, Sudijono (2011) stated that analyzing the distractors is aimed not only to know which items that cannot work properly but also to check why particular test taker failed to answer certain items

correctly. Arikunto (2005) also asserted that the distractor is effective if it has been chosen at least 5% total number of test takers, less effective if it is chosen less than 5% of the test takers, and ineffective if the distractor attracts more test takers from the upper group than those in the lower group. The result of the analysis showed that from total 160 distractors, 41% were functioned effectively, while 51% distractors were considered less effective which needed to be revised to function effectively, and 8% distractors were ineffective which needed to be replaced because they attracted more test takers from the upper group.

In line with the previous related studies that common mistakes in terms of test item quality, index of difficulty, discrimination power and effectiveness of distractors also discovered in this study. It can be summarized that theateacher-made English summativestest of the eleventh-grade students of the sample school was not appropriate to be used as a tool to evaluate the students' English learning because the test items did not fulfill all aspects of a good test in terms of quality, index of difficulty, discrimination power, and effectiveness of distractors.

CONLUSION AND SUGGESTION

Findings of this study proved that the quality of teacher-madesEnglish summative testsofsthe eleventh-gradesstudents of one public senior high school in Palembang was not a good test since 34 out of 40 items needed to be revised and replaced. The result of item analysis towards index of difficulty, discriminating power, and the effectiveness of distractors showed the test weaknesses. It was apparent that many of the items did not meet the satisfactory criteria in index of difficulty, discriminating power, and effectiveness of distractors. It can be concluded that the teachers of the school lacked of ability in constructing the test that was appropriate and suitable for the specific level.

In relation to the conclusion, some suggestions are offered for the teachers and school as an institution. For the teacher, it is necessary to have a good test items to measure students learning in English before conducting an assessment. Fulfilling all the criteria of a good test is a must. In this case, teachers are recommended to have good assessment literacy in test construction. For the school, it is recommended that the school provide teachers with sufficient supports in terms of professional development, such as in-house training or workshop dealing with test item construction so that teachers are literate and are able to construct good test.

REFERENCES

Ahmann, S. J., & Glock, M. D. (1967). *Evaluation growth principle of test and measurement*.

Arikunto, S. (2013). *Dasar-Dasar Evaluasi Pendidikan (Edisi Revisi)*. Jakarta: Bumi Aksara.

BSNP. (2010) *Panduan Penulisan Butir Soal*. Direktorat Pembinaan SMP Ditjen Menejemen Pendidikan Dasar dan Menengah Kementerian Pendidikan Nasional.

Djiwandono, S. (2011). *Tes Bahasa Pegangan Bagi Pengajar Bahasa*. Second Edition. Jakarta: PT Indeks.

Downie, N. M., & Heath, R. W. (1974). *Basic statistical methods.*

Ebel, R. I., & Frisbie, D. A. (1986). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prenctice-Hall.

Ghofur, A. (2004). *Pedoman Umum Pengembangan Penilaian Kurikulum Berbasis Kompetensi Sekolah Menengah Atas (SMA).* Jakarta: Depdiknas.

Gronlund, N. E., & Linn, R. L. (1965). *Measurement and evaluation in teaching* (Vol. 4). New York: Macmillan.

Henning, G. (2012). Twenty Common Testing Mistakes for EFL Teachers to Avoid. In *English Teaching Forum* (Vol. 50, No. 3, p. 33). US Department of State. Bureau of Educational and Cultural Affairs, Office of English Language Programs, SA-5, 2200 C Street NW 4th Floor, Washington, DC 20037.

Husna, H. H. (2012). *An Analysis of English Summative Test for 6th Grade Students in Three Public Elementary Schools in Udanawu District, Blitar Regency*.

Kristiana (2014). *An Analysis on the Content Validity of Summative Test for the Second Grade Students of Junior High School.*

Kurebwa & Nyaruwata (2013) *Assessment Challenges in the Primary schools: A Case of Gweru Urban Schools.*

Lahrichi, Anas. (2019). *Study on the Effectiveness of Formative and Summative Assessment Techniques in Education.*

Mardapi, D. (2008). *Teknik penyusunan instrumen tes dan nontes*.

Moore, K. D. (2014). *Effective instructional strategies: From theory to practice*. Sage Publications.

Simsek, A. (2016). A Comparative Analysis

of Common Mistakes in Achievement Tests Prepared by School Teachers and Corporate Trainers. *European Journal of Science and Mathematics Education*, *4*(4), 477-489.

Sudijono, A. (2011). *Pengantar Evaluasi Pendidikan.* Jakarta: Rajawali Pers

Thompson, B., & Levitov, J. E. (1985). Using Microcomputers to Score and Evaluate Items. *Collegiate Microcomputer*, *3*(2), 163-68.