English Journal
Vol. 12, No. 2; September 2018, pp. 62-67

TEACHER'S ASSESSMENT LITERACY: AN ANALYSIS OF MULTIPLE CHOICE TEST QUESTIONS OF A JUNIOR HIGH SCHOOL'S EFL LEARNERS

Muhammad Shabir
English Education Program
Faculty of Teacher Training and Education
Universitas Ibn Khaldun Bogor
m.shabir@uika-bogor.ac.id[1]

ABSTRACT

Stick to the required rules and standards of multiple choice test assessment, the study seeks to analyze the multiple choice test questions administered for Indonesian EFL learners at a junior high school. A careful analysis was given to the items of a summative multiple choice of English midterm test. Using the classical test theory (CTT) in analyzing each test item, it was found that only 23 percent of the total of the analyzed items are acceptable or have adopted the standards or the rules required. There are two identified major problems or causes to the poor quality of the test: distractor plausibility and limitedness in the number of possible options or distractors. To deal with the problems, the study suggests the using of fewer possible options or alternatives for such test for better assessment.

Keywords*:* Multiple choice test, EFL learners, assessment, CTT

INTRODUCTION

The notion of assessment is central and fundamental when it refers to the evaluation of teaching and learning practices. For this reason, experts have invested and will always devote their thought and energy to find out what might be the most appropriate and functional method for such purpose (McMillan, 1997). Basically, assessment refers to any method, strategy, or tool that a teacher may use to collect evidence about students' progress toward the achievement of established goals. In other words, it is a process of collecting information and gathering evidence about what students have learned. The problem is that it is not an easy task to make a good assessment (Kibble, 2016). Oftentimes, the assessment used does not reflect the actual learning situation (Heaton, 1990). This suggests that assessment should actually be aimed at evaluating the strengths and weaknesses of students' learning (Popham, 1995). To achieve this goal, assessment must be constructed properly and meaningfully.

Inspired by theoretical enclaves above, this paper tries to analyze and observe a summative test administered for a junior high school's EFL students, with a guiding question to address: "Do the constructed multiple choice test questions for a junior high school's EFL students reflect the standards or the rules required?"

The test was taken by 110 eleventh grade students which consisted of 40 items with a total of 40 answer keys and 120 distractors. Anchored to the multiple choice test rules *(Haladyna, T. M., Downing, S. M., & Rodriguez, M. C., 2002)*, *each* item will be analyzed using the Classical Test Theory/ CTT (Lynch, 2003). The rules include distractor possibility, using plausible distractors, using a question format, emphasizing higher-level thinking, keeping similarity in the length of option, using correct grammar, avoiding clues to the correct answer, avoiding negative questions , using only one correct option, giving clear instructions , using only a single clearly-defined problem and including

the main idea in the question, avoiding "none of the above" option, avoiding using questions when distractors are limited or assessing problem-solving and creativity. Guided by these rules, each item will weighed by computing *item difficulty* and analyzing *distracter/incorrect alternative*.

Inspired by theoretical enclaves above, this paper tries to analyze and observe a summative test administered for a junior high school's EFL students, with a guiding question to address: "Do the constructed multiple choice test questions for a junior high school's EFL students reflect the standards or the rules required?" The study aims to address whether the constructed multiple choice test questions for a junior high school's EFL students reflect the standards or the rules required. The result of the study could be useful for those who have the authority to enhance the understanding of teachers of how to construct an effective test in evaluating students' learning.

LITERATURE REVIEW

The notion of assessment is central and fundamental when it refers to the evaluation of teaching and learning practices. For this reason, experts have invested and will always devote their thought and energy to find out what might be the most appropriate and functional method for such purpose (McMillan, 1997). Basically, assessment refers to any method, strategy, or tool that a teacher may use to collect evidence about students' progress toward the achievement of established goals. Assessment is the process of gathering data. More specifically, assessment is the ways instructors gather data about their teaching and their students' learning (Hanna & Dettmer, 2004). In other words, it is a process of collecting information and gathering evidence about what students have learned. The problem is that it is not an easy task to make a good assessment (Kibble, 2016). Oftentimes, the assessment used does not reflect the actual learning situation (Heaton, 1990). This suggests that assessment should actually be aimed at evaluating the strengths and weaknesses of students' learning (Popham, 1995). To achieve this goal, assessment must be constructed properly and meaningfully.

Formative and Summative Assessment

Generally, there are two types of assessment: formative and summative. As to formative type, it is typically not graded and act as a gauge to students' learning progress and to determine teaching effectiveness (Hanna & Dettmer, 2004). This assessment is used to identify areas that may need improvement. Hanna and Dettmer (2004) suggest that formative assessment provides feedback and information during the instructional process, while learning is taking place, and while learning is occurring. In other words, formative assessment measures student progress but it can also assess teacher's own teaching progress.

As to summative assessment, this type of assessment takes place once the learning has been concluded. This aims to provide teachers information on how well the teaching and learning process have been carried out. At this stage, formal learning is no more conducted. Hanna and Dettmer (2004) suggest that in summative assessment, teachers should develop around a set of standards or expectations so that students understand what is expected of them for each of the criteria.

As widely implemented, summative assessments are administered when students have completed their studies or at the end of the semester. This assessment is to evaluate what they have learned and how well they learned. Hanna and Dettmer (2004) say that grades are usually an outcome of summative assessment: they indicate whether the student has an acceptable

level of knowledge-gain. Through this evaluation, teachers will be able to find out whether the students are able to effectively progress to the next part of the class or to the next course in the curriculum or to the next level of academic standing. To this far, it is clear that summative assessment is more product-oriented and assesses the final product.

Multiple Choice Questions: a common type of assessment

Multiple choice questions is a type of assessment which is widely used in evaluating students' performance. However, it is not easy task to construct good test items, it requires a good knowledge of the content and understanding of the objectives of assessment as well as good skills in writing the items (Walsh K. 2005). Normally, multiple choice question uses four or even five options. However, this format also can be reduced to three by maintaining the quality of the test. Studies by Grier (1975) show that multiple choice question with three options could increase reliability of the test. Green et al. (1982) also showed that three-option multiple choice question could improve validity of a test. Haladyna and Downing (1985) in their review of research on multiple choice question showed mixed results for item discrimination. In their review, while one study showed no difference in item discrimination between three and four options, another study showed three-option items to have better item discrimination than four options. However, later studies showed an increase in item discrimination with three-option. In terms of item difficulty, Haladyna and Downing reviewed studies on the number of options in terms of item difficulty and concluded that three-four options are optimal. In their observation, they took into account the issue of guessing which is more common for low performers. They concluded that for most examinees three-options appeared to be optimal. A comparison of three-and four option items showed a decrease in 'test-wiseness' or guessing with three-option items. 'Test-wiseness' was defined by Millman et al. as 'a subject's capacity to utilize the characteristics and formats of the test and/or test taking situation to receive a high score'. As to the useful options, Haladyna and Downing concluded that the 3-option format is optimal as the number of functional distracters per item was optimal. Other studies confirmed that the three-option format had fewer dysfunctional distractors, the mean number of functioning distractors was much lower than two and reducing the least popular option had only a minimal effect on the performance of the remaining options.

METHOD

Data of this descriptive study composed of one summative multiple-choice test of English subject along with a total of 110 answer sheets from three classes of a junior high school's students. These answer sheets were part of the students' 1st semester midterm tests administered in 2018. The test consisted of 40 items with four options: one correct answer and three distractors. The analyses started by calculating item difficulty *p value* (the proportion between the examinees with the correct answer and those with the incorrect answer). The p-value was calculated by $p = [(H+L) / N] \times 100$. N is the total number of students in both high and low groups. H and L are the number of correct responses in the high and low groups, respectively. Items with p-value between 30 - 70 were considered as acceptable (Mozaffer R.H., Farhan J (2012). Each correct response was awarded 1 mark. Thus, the maximum possible score of the overall test was 40 and the minimum 0. This then followed by observing *distracters/incorrect alternative* of each item. A particular

attention was given to the undesirable or unacceptable distracters and the confirming reasons to such situation.

FINDINGS AND DISCUSSION

At this part, the findings and relation to the acceptability of test item will be described and analyzed.

The following tables show the students' answers to a multiple choice test with a total of 40 questions. Each question has four options: A,B,C, and D. The *marked * column* denotes the correct answers and number of students with this option.

Question	A	B	C	D
1	110*	0	0	0
2	110*	0	0	0
3	110*	0	0	0
4	0	0	0	110*
5	0	0	110*	0
6	110*	0	0	0
7	0	110*	0	0
8	110*	0	0	0
9	0	0	110*	0
10	110*	0	0	0
11	0	110*	0	0
12	1	0	109	0
13	110	0	0	0
14	5	15	90*	0
15	0	0	110*	0
16	109*	0	0	1
17	1	1	107*	1
18	110*	0	0	0
19	0	110*	0	0
20	0	118*	1	1

Question	A	B	C	D
21	0	110*	0	0
22	110*	0	0	0
23	1	0	109*	0
24	1	1	107*	1
25	1	107*	1	1
26	109*	0	1	0
27	95*	15	0	0
28	0	110*	0	0
29	80*	5	10	15
30	0	0	110*	0
31	107	1	1	1
32	0	0	0	110*
33	3	101*	2	4
34	2	1	107*	0
35	0	0	0	110*
36	110*	0	0	0
37	104*	6	0	0
38	0	0	0	110*
39	1	1	108*	0
40	0	0	0	110*

Based on the tables, there are as many as 25 questions or 62.5 percent of the total of the answer sheets with no selected distractors. All these items' distractors evidently failed to attract any student. The item difficulty calculated by dividing the number of students who choose the correct answer by whole number of students also points out that those items are very easy since the P-value for each of those 25 questions is above 0.90. All of the students could easily answer these items correctly with no any selected distractor. The study tells that all those provided distractors were not functional or none could attract students' attention. It can be concluded that those items were not constructed based on the required rules, and therefore, are not worth testing. There was an obvious distressing fact to the poor quality of the items after a careful analysis was conducted. Most of the questions are visibly leading the students to the clue of which of the provided options is the most possible answer. This is contradictory to the required practices in constructing multiple choice questions test in which the clue to the correct answer should be avoided (Haladyna et al. 2002).

The table also shows not all distractors in one item which are selected by students. There are six questions or 15 percent of the total of items categorized into this type, which are 14, 17, 20, 26, 34, and 39. Having analyzed those five questions, the problem to the unselected distractors relies on the impossibility of option. Impossible options were often found that left them unselected by students. Presumably, those questions have limited possible answers and are not fit for multiple choice questions test format. As it is suggested that answer options in multiple choice question should be plausible and corresponding to the students' real understanding (Haladyna et al. 2002). Other items, excluding those six and 25 easy items discussed earlier, only can be categorized as the items with acceptable or desirable distractors.

Based on the findings, the study concludes that there are two main factors of the poor quality of the test. The first is related to the plausibility of distractors.

Many seem not possible and are often beyond students' real understanding. The second relies on the availability of possible options. This means some distractors provided seem exaggerated or beyond the context. To deal with the problems, the study suggests the use of fewer options or alternatives for such test for better assessment.

CONCLUSION

The study has shown the fact of teachers' literacy in constructing multiple choice test items at a school in Bogor. The poor quality of the test confirms the less understanding of the teachers of how to construct good test items which can effectively and accurately measure students' learning performance. Based on the study, plausibility and number of distractors are two factors that must be considered. The first is related to plausibility. Many test items are not possible and often beyond students' real understanding. This contradicts the goal of the test which should be used to evaluate students' performance. The second relies on the availability of possible options. The study has indicated that a test item which is moderately difficult is not suggested to have four options. For this case, fewer options could be solution in order to improve validity of the test.

The conclude, the study shows an indication that teacher's literacy in making assessment is not in line with the accepted procedure of how a test should be administered and given in appropriate context. This research is useful since it can be important information for the responsible institutions whose authority is to enhance better the understanding and literacy of teachers how a test should be constructed in an effective and appropriate manner. This research report has unboxed the wide practices of improperly constructed test item which consequently may fade the students' real performance in their learning.

REFERENCES

Haladyna T.M., Steven M. Downing & Michael C. Rodriguez (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment, Applied Measurement in Education, 15:3, 309-333.

Hanna, G. S., & Dettmer, P. A. (2004). Assessment for effective teaching: Using context-adaptive planning. Boston, MA: Pearson A&B.

Heaton, J. (1990). Writing English language tests. New York: Longman Inc.

Hingorjo RH., Jaleel F (2012). Analysis of one-best MCQs: the difficulty index, discrimination index and distractor efficiency. J Pakistan Med Asso. 2012 Feb 1;62(2):142.

Kibble, J. D. (2016). Best practices in summative assessment. Adv Physiol Educ 41: 110–119.

McMillan, J.H. (1997). Classroom Assessment: Principles and Practices for Effective Instruction. Boston: Allyn & Bacon.

Mozaffer Rahim Hingorjo, Farhan Jaleel (2012). Analysis of OneBest MCQs: the Difficulty Index, Discrimination Index and Distracter Efficiency. Journal of Pakistan Medical Association.

Popham, W.J. (1995). Classroom assessment. Needham Heights, MA: Allyn & Bacon.

Walsh K. (2005). Advice on writing multiple choice questions (MCQs). BMJ Careers 2005. Available at http://careers.bmj.com/careers/advice/view-article.html?id=616 (accessed on 23 May 2019.

Millman, J., Bishop, H.I., & Ebel, R. (1965). An analysis of test wiseness. Educational and

Psychological Measurement, 25(1), 707-72