ANALYSING A MULTIPLE-CHOICE TEST TO FIND OUT THE ITEM FACILITY (IF), THE ITEM DISCRIMINATION (ID) AND THE DISTRACTOR EFFICIENCY

Sri Ariani[1], Rini Hendrita[2]
Universitas Muhammadiyah Sumatera Barat
email: sri.ariani80@gmail.com

Abstrak

Makalah ini bertujuan untuk menganalisa sebuah perangkat test atau ujian yang populer digunakan sebagai penilaian kelas terutama di kelas berukuran besar yaitu tes pilihan ganda. Tes pilihan ganda yang akan dianalisis dalam makalah ini adalah tes yang digunakan pada ujian akhir semester bahasa Inggris untuk siswa SMA kelas sebelas yang digunakan pada semester pertama. Tes ini dibuat oleh tim guru yang mengajar di lima SMA negeri di satu kota di Indonesia dan telah digunakan sebagai ujian akhir untuk lima sekolah ini. Penelitian ini menggunakan metode kuantitatif untuk mengetahui Item Facility (IF), Item Discrimination (ID) dan Distractor Efficiency dari tes tersebut. Setelah melihat pembedaan butir soal tes ini, ternyata ada beberapa butir soal yang menunjukkan ketidakjelasan atau bahkan jawaban yang dinyatakan oleh guru tidak benar. Ini dapat mempertanyakan validitas soal tes. Oleh karena itu tes pilihan ganda sebaiknya tidak digunakan sebagai alat tunggal untuk menilai kemampuan siswa karena hasil yang diberikan kurang valid, otentik, dan memberikan umpan balik yang tepat bagi guru dimana hal ini merupakan prinsip terpenting dalam melakukan penilaian.

Kata Kunci*: item facility, item discrimination and the distractor efficiency*

Abstract

This paper is going to analyse a subset of assessment which is popularly used as classroom assessment especially in the large size classroom that is multiple-choice test. The multiple-choice test that will be analysed in this paper is an English final test for Senior High School Students eleventh grade used at first semester. This test was made by teachers' team teaching of five state senior high schools in one city in Indonesia and had been used as a final test for these five schools for their eleventh-grade students at their first semester English final test. This research uses quantitative method in order to find out Item Facility (IF), Item Discrimination (ID) and Distractor Efficiency of the test. After looking at the item discriminations of this test, it reveals that some of the test items show ambiguity or even the answers stated by the teacher are not correct. These can question the validity of the test item. Therefore multiple-choice test should not be done as a single tool to evaluate the ability of the students because it is lack validity, authenticity, and wash back which are the most important principles in conducting the assessment.

Keywords: item facility, item discrimination and the distractor efficiency

INTRODUCTION

Teaching is not only the main concern of a teacher. Teacher is also responsible to monitor progress of the students and evaluate whether the students have achieved the standards at the end of a learning sequence or a unit. Therefore, the teacher needs to assess the students. However, classroom assessment should be a combination of formative and summative assessments. Formative assessment serves as feedback for the students to develop their learning while summative assessment is useful to summarize students' achievement at the end of unit instruction to see how well they have reached the goal. In order to fulfil its functions, the assessment itself should be evaluated.

Therefore, this paper is going to analyse a subset of assessment which is popularly used as classroom assessment especially in the large size classroom that is multiple-choice test. The multiple-choice test that will be analysed in this paper is an English final test for Senior High School Students eleventh grade used at first semester. This test was made by teachers' team teaching of five state senior high schools in one city in Indonesia and had been used as a final test for these five schools for their eleventh-grade students at their first semester English final test. Though the test has been used several years ago, this paper is going to show how the teachers or instructors analyse whether their test has met the need of assessment.

As a reference for evaluating the validity of the test especially content validity, the writer uses the English syllabus used for eleventh grade students. Furthermore, from the population of all eleventh-grade students of this state senior high school, the writer takes one class as a sample to evaluate the items in the multiple-choice test. The sample class is an eleventh grade of science class which consists of 31 (thirty one) students/ test takers. The score of these students are then used to evaluate the test items in this multiple-choice test in order to find out the Item Facility (IF), the Item Discrimination (ID) and the Distractor efficiency. The calculation of its items facility, item discrimination and distractor efficiency are provided at appendix three, four and five.

Therefore, this paper will analyse whether this multiple choice is practical, reliable, valid, authentic and provides wash back for the students. And it also analyses level of difficulty of each test items, determines whether its test items can distinguish between the low ability students and the high ability students. This paper also analyse the efficiency of the test distractors and the distribution of the correct response among the distractors.

REVIEW OF RELATED THEORIES

Test as a part of assessment is a tool to evaluate the achievement of the students. Chandio and Jafferi (2015) say that it is an instrument which consists of techniques, procedures or items to measure the test takers' performance. Brown and Priyanvada (2019) advocate their statement and defines test as a kind of assessment technique which is prepared administratively, the time is identified in the curriculum and the students know that they should give their peak performance to be measured.

1. Principles of language assessment

As an instrument, it must fulfil the qualification of a good instrument. Now, we come to the question to determine whether a test has been good or not. Brown and Priyanvada (2019) give some questions to determine whether a test is good or not. First we should ask whether the test can be given within appropriate administrative constraints. Then, the test should accurately measure what we want it to test. Thirdly, we should ask whether the language in the test represents the real-world language use. The test should also provide useful information for the learners. Brown says that these five questions can be the criteria for testing a test; they are called practicality, reliability, validity, authenticity and wash back.

1.1 Practicality

Now, we come to the question of how we know that a test is practice. Bachman and Palmer (2009) define that practicality is a condition in which the test is reasonable or logical to be made, given and scored. In order to meet this condition, the test should fulfil some criteria.

The following are criteria of practicality proposed by Hughes and Jake (2020). First the cost to administer and also design the test should be logical. Then the time should also reasonable. The time is not only limited to the test administration, but it should also consider the time for the examiner to evaluate test takers' results. The direction of the test should be clear for the test takers and the test administers. Clear direction is not only on the test instruction, but also on the scoring procedures. The available human resources should also be able to utilise the test. Thus, the test that totally depends on computer is impractical if the test takes place away from the nearest computer. The test then should also not exceed the available material resources. And the last, the effort to designer and score the test should also be considered.

1.2 Reliability

Test does not only have to be practical but also reliable. Now, we question how we know that the test is reliable. Hughes and Jake (2020) give the key terms for reliable; they are consistent and dependable. The test is called consistent and dependable if the score obtained from conducting the same test towards the same or matched students at two different occasions are similar.

According to Brown and Priyanvada (2019), there are four factors which can influence the reliability of the test. First factor is students' physical or psychological conditions such as illness, fatigue and anxiety. This factor belongs to students-related reliability. The second factor is related to the rater reliability. It can be inter-rater reliability or intra-rater reliability. Inter-rater reliability can be reached if the scores given by two or more scorers towards the same test are consistent. Whereas, intra-rater reliability is reached if a scorer rates consistently towards all students without influenced by unclear criteria, fatigue or bias. In multiple choice test, rater reliability can be reached if the fix responses have been determined in advance. In subjective test with open-ended questions, the rater reliability can be increased by using clear analytical scoring instruments.

Besides student-related and rater reliability, there are also test administration reliability and test reliability. Test administration reliability is influenced by the condition of test administration while test reliability is influenced by the test itself such as the distribution of correct answer in objective test, rater bias in determining correct answer, printed item clarity and the length of the test.

1.3 Validity

The third criterion of a good test is valid. According to Cyril in Clark (2020), a test is called valid if the result of the test is appropriate, meaningful and useful in assessing students' ability. A valid test also measures exactly what it proposes to measure, samples test’s criterion or objectives and based on empirical evidences.

Lou and Jiayu (2021) explain that there are four types of empirical evidences. The first one is content-related evidence in which the test tests the subject have been taught. Test taker should also directly perform what should be measured or in other word direct testing. For example, if the test wants to test students’ ability in speaking, the test should ask the students to perform something orally.

The second empirical evidence is criterion-related evidence. This evidence shows the extent to which the students has reached the criterion or objectives that have been specified in advance. And the third evidence is construct-related evidence which shows to what extent the test tap into the defined theoretical constructs.

Next evidence is consequential validity or the impacts given by the test such as the effect on the preparation of the test-takers and social interpretation of the test or the use of the test. And the last evidence is face validity that is the extent to which the students view the test is fair, relevant and useful to improve learning. The relevance can be reached if the students see that the test is well-constructed, familiar in the format, relevant in the time setting, has clear and uncomplicated items, has clear direction and relates to the course work and reasonable in the level of difficulty.

1.4 Authenticity

Another criterion of a good test is authentic. Malley, Michael & Valdez (1996), Marheini at all (2014) and Reynisdottir (2014) states that an authentic test is the samples the real world. It samples the real worlds if the language used is natural, the item is contextualized, the topics are meaningful, relevant and interesting, the test is thematic replicate real-world tasks.

1.5 Wash back

The last criterion is wash back which means the extent to which the test affects teaching and learning. Ojung & Allida (2017) state that there are several ways to know whether the test affects teaching and learning. First, we should see whether the test give positive influence on the way teacher teach and the way students learn. Then the test should offer learners a chance to prepare in order to give peak performance. The feedback given by the test should also enhance the students.

2. Multiple-choice Items Validity

Multiple-choice test seems to be the most popular test for large scale test. It is also often used as a classroom language test in which the number of test takers is large. In order to design an appropriate multiple-choice test, there are several things that should be considered by the test designer.

Hughes and Jake (2020) propose several guidelines in designing multiple-choice items. First, each item should measure a single objective. Then, the stem (i.e. the body of the test item) and the options (i.e. the alternatives to be chosen) should be stated as simply and directly as possible. Thus, needless redundancy should be removed. Third, the intended answer is clearly only one. Last, three item indices should be considered to accept, revise or even discard the items. The three indices are item facility, item discrimination and distractor efficiency.

METHODOLOGY

This research uses quantitative method in order to find out Item Facility (IF), Item Discrimination (ID) and Distractor Efficiency of the test.

Item Facility (IF)

Item facility shows the level of item’s difficulty. Brown and Priyanvada (2019) state that the item achieves validity if it can successfully separate between the low ability students with the high ability students. It will not be reached if the item is too easy that almost all students response correctly or if the item is too difficult so that mostly all student give wrong answer.

Brown and Priyanvada (2019) give the following formula to find the item facility:

$$\boxed{\mathrm{IF} = \frac{\text{number of correct answer for an item}}{\text{number of students respond that item}}}$$

An appropriate test items should have IFs ranging from 0.15 which means very difficult to 0.85 which means very easy. Giving occasional very easy items has two advantages, firstly it can give a feeling of success among low ability students and secondly it can be a warm-up item for them. While very difficult items challenge the highest ability students.

Item Discrimination (ID)

Item discrimination has a function to show how well the items discriminate between high ability students and low ability students. All students’ scores first are ranked from highest to lowest score. This rank- ordered scores are then divided by three: the highest score group, the middle score group and the lowest score group. The middle group is then eliminated, thus only the answers of the high and low rank group will be used to get item discrimination of each item.

The formula to count item discrimination as proposed by Brown and Priyanvada (2019) is as follows:

$$\boxed{\mathrm{ID} = \frac{\text{high group no. correct} - \text{low group no. correct}}{\frac{1}{2}\ \text{total of two comparison group}}}$$

A perfect high discriminating power is 1.0 and no discriminating power at all is shown by ID zero. Thus, the practical use of ID is to select items from a test bank to be included in the test or to discard or improve some items with lower ID.

Distractor Efficiency

Distractor efficiency is useful to see the efficiency of the distractors to lure the low ability students and to see the distribution of the responses across all distractors. This distractor efficiency is looked at the choices of answers from the high ability group and low ability group.

RESULT AND DISCUSSION

The writer evaluates the test items based on the criteria proposed by Brown as been stated in the review of related literature.

1. **Item Facility (IF)**

From the number of the students who answers each items correctly which is divided with the number of students who response each item, we can get the difficulty level of each items.

The following table shows the item facility of each item which is got from the formula given by Brown. List of students who answer correctly for each item are shown in appendix 3.

Table 1. Item facility of each item

No	1	2	3	4	5	6	7	8	9	10
IF	1	0.93	0.61	0.03	0.90	0.54	0.67	0.74	1	0.96
action	revised	revised		revised	revised				revised	revised

No	11	12	13	14	15	16	17	18	19	20
IF	0.94	1	1	1	0.55	0.94	0.94	0.87	1	0.81
action	revised	revised	revised	revised		revised	revised	revised	revised	

No	21	22	23	24	25	26	27	28	29	30
IF	1	1	0.25	0.67	0.74	0.22	0.93	0.51	0.74	0.61
action	revised	revised					revised			

No	31	32	33	34	35	36	37	38	39	40
IF	0.23	1	0.65	0.42	1	0.03	0.65	0.77	0.94	0.65
action		revised			revised	revised			revised	

No	41	42	43	44	45	46	47	48	49	50
IF	0.16	0.48	0.74	0	0.97	0.61	0.65	0.48	0.97	0.84
action				error	revised				revised	

From the table we can see that there are 10 (ten) numbers with yellow high light whose item facility is 1 (one). It means these items are too easy and fails to discriminate between the high ability students and low ability students. The number of other items whose item facility is more than 0.85 which means quite easy is 11 (eleven) numbers highlighted in grey color. Whereas the number of items whose item facility is less than 0.15 which means too difficult is 2 (two) numbers typed in red color. Total number of item facility which is out of the range suggested by Brown which is between 0.15 to 0.85 is 23 (twenty) items.

As being suggested by Brown, these items especially for the too difficult items should be revised because they are highly beyond the range of suggested difficulty level. The last 21 numbers which are too easy may be revised half and kept half in order to give the feeling of success for the low ability level of students.

2. Item Discrimination

Item discriminations for each test items got from this class are shown in the appendix 4. The following table is the result of item discriminations in appendix 4.

Table 2. Item Discriminations of each item

No	1	2	3	4	5	6	7	8	9	10
ID	0	0	0	0	0	0.4	0.5	0.6	0	0.1

No	11	12	13	14	15	16	17	18	19	20
ID	0.2	0	0	0	0.2	0	-0.2	0.1	0	-0.2

No	21	22	23	24	25	26	27	28	29	30
ID	0	0	0.2	-0.1	0.2	-0.2	0.1	0.2	0.5	-0.1

No	31	32	33	34	35	36	37	38	39	40
ID	-0.1	0	0.7	0.6	0	-0.1	0.1	0.1	0.2	-0.3

No	41	42	43	44	45	46	47	48	49	50
ID	0.3	0.3	0.2	0	0.1	0	0.3	0	0.1	0.1

The table shows that there are 18 (eighteen) numbers highlighted in yellow which cannot discriminate between high level ability with low level ability students. And it is more astonishing that there are some numbers which are more answered correctly by the low level ability students so that the item discrimination becomes negative. They are typed in red colour. Next part will reveal the reasons why these items discrimination can be negative by seeing the distractor efficiency of each test items.

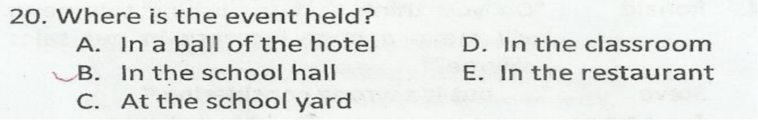
3. Distractor Efficiency

Table 2 shows that some items discrimination are negative. Let we discuss some of them by looking at its distractor efficiency at table 3.

Table 3. Distractor efficiency for number 20 whose item discrimination is -0.2.

Item	Choices	A	B	**C**	D	E
20	High-ability students 10	0	2	7	0	1
	Low-ability students 10	0	0	9	0	1

The correct answer is C. The text can be seen in the appendix 1.



20. Where is the event held?

A. In a ball of the hotel D. In the classroom

[✓ B.] In the school hall E. In the restaurant

C. At the school yard

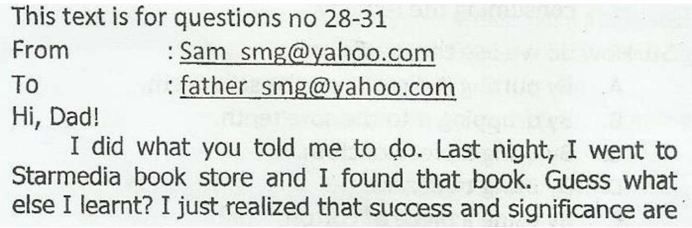
The distractors show ambiguity because the text does not state the location of the event. Therefore the answer can be B and C. Therefore, there are 2 high ability students who chose B. Therefore, this item is not valid because it does not fulfil the requirement of a good test item suggested by Brown that there should only one correct answer for each item.

Another item which shows ambiguity is test item number 31. Table 4 shows the distractor efficiency of the item.

Table 4. Distractor efficiency for number 31 whose item discrimination is -0.1.

Item	Choices	**A**	B	C	D	E
31	High-ability students 10	2	4	0	4	0
	Low-ability students 10	4	3	0	3	0

The correct answer is A. The following is the item of the test.



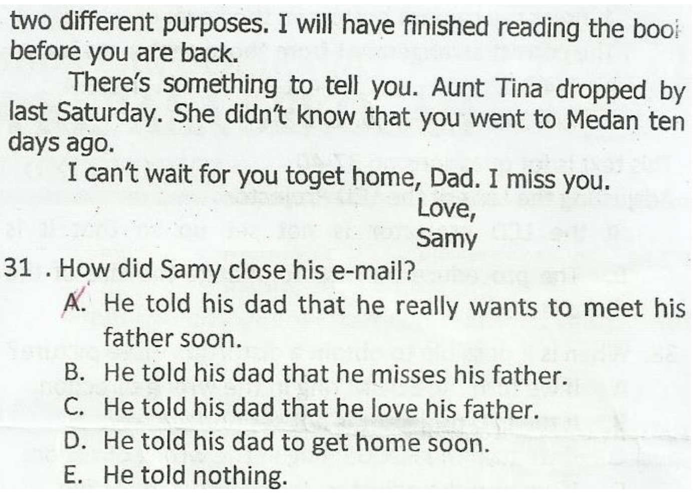
This text is for questions no 28-31

From : Sam_smg@yahoo.com

To : father_smg@yahoo.com

Hi, Dad!

I did what you told me to do. Last night, I went to Starmedia book store and I found that book. Guess what else I learnt? I just realized that success and significance are



two different purposes. I will have finished reading the boo
before you are back.

There's something to tell you. Aunt Tina dropped by last Saturday. She didn't know that you went to Medan ten days ago.

I can't wait for you toget home, Dad. I miss you.

Love,
Samy

31 . How did Samy close his e-mail?

A. He told his dad that he really wants to meet his father soon.
B. He told his dad that he misses his father.
C. He told his dad that he love his father.
D. He told his dad to get home soon.
E. He told nothing.

The correct answer should be D because this item tests about indirect sentence. The item A which is determined as the right answer by the teacher is absolutely false because it should be "He told his dad that he really wanted to meet his father soon. If the main clause uses simple past tense, than the dependent clause should also be changed into past form. The answers of the students shows that four high ability students answer D.

CONCLUSION

After looking at the item discriminations of this test, it reveals that some of the test items show ambiguity or even the answers stated by the teacher are not correct. These can question the validity of the test item. Therefore, making multiple-choice is not as easy as selecting the available test items from the bank of tests. After the items have been used, the teacher as the test administrator should count the validity of the test items based on the answers of the test-takers. This calculation can be used as the consideration for the next test administration whether the items will be used, revised or even discarded.

Moreover, multiple-choice test should not be done as a single tool to evaluate the ability of the students because it is lack validity, authenticity, and wash back which are the most important principles in conducting the assessment. Teacher should combine formative assessment as their daily classroom assessment to develop students' ability and use another possibly of performance based assessment which really provides valid measurement towards students' competencies.

References

Bachman, Lyly F & Palmer, A.S. 2009. Language Testing in Practice: Designing and Developing useful Language tests. New York. Oxford University Press

Brown, H. Douglas & Priyanvada Abeywickrama. 2019. *Language Assessment, Principles and Classroom Practice, third edition*. New York: Pearson Education

Chandio, Muhammad Tufail & Saima Jafferi. 2015. '*Teaching English as a Language not Subject by Employing Formative Assessment' Journal of Education and Educational Development Vol. 2 No.2*

Clark, Tony. 2020. Lessons and Legacy: A Tribute to Professor Cyril J Weir. Cambridge: Cambridge University Press.

Cyril J. Weir. 1990. *Communicative Language Testing*. London. Prentice Hal.

Hughes, Arthur & Jake Hughes. 2020. *Testing for Language Teacher, third edition*. Cambridge: Cambridge University Press.

Lou, Kaizhou & Jiayu Wang. 2021. Validity Arguments in Language Testing. Case Studies of Validation Research. Language Assessment Quarterly. Cambridge: Cambridge University Press.

Marhaeni. A.A.I.N et all. 2014. Toward Authentic Language Assessment: A Case in Indonesia EFL Classroom. The European Conference on Language Leaning.

Ojung. J & Allida. D. 2017. A Survey of Authentic Assessment Used to Evaluate English Language Learning in Nandi Central Sub-County Secondary Schools, Kenya. Baraton Interdicipline Research Journal, 7 (special issue)

O’ Malley, J. Michael & Pierce, Lorraine Valdez.1996. Authentic Assessment for English Langauge Leaners: Practical Approaches for Teachers. Addison-Wesley Publishing Company.

Reynisdottir. B.B. 2016. The Efficacy of Authentic Assessment. University of Iceland