QUALITY OF POSTTEST ITEMS ADMINISTERED BY LANGUAGE CENTER (PUSBA) IAIN RADEN INTAN LAMPUNG

M. SAYID WIJAYA
sayidwijaya@gmail.com
IAIN Raden Intan Lampung

Abstract: In achieving learning objectives, it needs a test to measure students' achievement. A test should be constructed based on the learning objectives so it measures what it really intends to measure. It also should be assured that it fulfills the criteria of good test, one of which is item validity. Item validity concerns on how each item measures what has been established in blueprint. However, analyzing item validity is insufficient since it is only a part of item analysis. Therefore, the urge to analyze item quality is imperative. Further, the posttest instrument used in measuring students' achievement after following matriculation program in IAIN Raden Intan Lampung has not been analyzed yet since the matriculation program operated. In this case, the quality of test items is crucial to be analyzed. In analyzing quality of test items, ANATES was employed to investigate item discrimination, item difficulty, and item validity for practicality and efficiency. The result demonstrated that posttest instrument needed to be revised in terms of item difficulty and effectiveness of distractors.

Key words: item analysis

In language teaching, at the end of a course or unit of instruction, we are concerned primarily with the extent to which the students have achieved the intended outcomes of the instruction (Gronlund and Waugh, 2009:9). In this case, the success of the teaching and learning process is determined by the achievement of instructional objectives. Those instructional objectives should be reflected by students' achievement which is investigated by administering certain tests designed to achieve those objectives. A test is designed to dig up students' insights. In simple terms, a test is a method of measuring a person's ability, knowledge, or performance in a given domain (Brown, 2004:3). In other words, to

find out whether or not specific competencies or instructional objectives have been achieved, it needs to measure or assess students' knowledge of a certain area.

The result of the test commonly becomes a consideration to evaluate the overall teaching and learning process occurs. The test used for that purposes is known as achievement test. This test is related directly to classroom lessons, units, or even a total curriculum. This achievement tests are (or should be) limited to particular material addressed in curriculum within a particular time frame and are offered after a course has focused on the objectives in questions (Brown 2004:47). It means that this test focuses on established objectives which have been set up in the beginning of the teaching and learning in classroom context and should be measured in the end of the language program. Thus, this test must be designed with very specific reference to a particular course (Brown, 1996:14).

In designing such a test, it needs very clear procedures which will help a test maker to keep the intended test being in line with its objectives. In this case, selecting an appropriate approach in language testing becomes a decisive factor to make sure that the aspects to be measured are all covered in the test. Two popular approaches in language testing are discrete point testing and integrative testing. Discrete point testing refers to the testing of one element at a time, item by item. This might, for example, take the form of a series of items, each testing a particular grammatical structure. Integrative testing, by contrast, requires the candidate to combine many language elements in the completion of a task. This

might involve writing a composition, making notes while listening to a lecture taking a dictation, or completing a cloze passage (Huges, 2003:19).

Besides determining an approach in language testing, a test specification is also needed to be a guideline in constructing the test. It provides the official statement about what the test tests and how it tests it. The specifications are the blueprint to be followed by the test and item writers (Alderson et.al., 1995:9). They provide details information of the test development so a test maker will keep on track of the test constructed. A set of specification for the test must be written at the outset. This will include information on: content, test structure, timing, medium/channel, techniques to be used, criteria levels of performance, and scoring procedure (Huges, 2003:59).

When a test specification has been established, test items are ready to be written. Test items are derived from the test specification by referring to the aspects measured and the approach employed. However, a test maker cannot merely guarantee that the test constructed is a well-made. In this case, the test should fulfill the criteria of good test, validity and reliability. A test is said to be valid if it measures accurately what it is intended to measure (Huges, 2003:16). It means that if the test is designed to measure students' performance in speaking, it must ask the students to speak. Further, there are two kinds of validity which are commonly regarded crucial, content validity and construct validity. Content validity depends on a careful analysis of the language being tested and of the particular course objectives. The test should be so constructed as to contain a

representative sample of the course, the relationship between the test items and the course objectives always being apparent (Heaton, 1988:160). In short, this kind of validity can be provided by displaying test specification which covers all aspects or sub aspects to be measured for each test item in accordance with the objective of the test. Different from content validity, a test has construct validity if it is capable of measuring certain specific characteristics in accordance with a theory of language behavior and learning. This type of validity assumes the existence of certain learning theories or constructs underlying the acquisition of abilities and skills (Heaton, 1988:160). In this case, making sure that the aspects or components of the skills or dimension to be measured in test specification is based on certain theories grants the test to have construct validity. When the aspects or components are not supported by relevant theories or are objected by another theory, it possibly causes the test has low construct validity.

Leaving aside validity of the test, another criterion for good test is reliability. Reliability refers to the consistency of test scores – that is, to how consistency they are from one measurement to another (Gronlund, 1977:138). In tests constructed of items that can be scored correct of incorrect, each item should provide additional information about the ability of a test taker on the construct in question. By ensuring that responses to individual items are not dependent upon the responses to other items, that they have good facility values and discrimination, and that we have enough items, we can ensure that such test have the quality of reliability (Flucher and Davidson, 2007:104). In other words, if the

test is administered repeatedly, it is likely possible the test will provide such a consistency in resulting the test scores.

In finding out whether or not a test has validity and reliability evidence, tryout is administered. Content validity and construct validity are demanded to be validated by the experts since they are the ones who are majoring the concern of the components of the test or the skills to be measured. Reliability of the test is computed for its reliability coefficient. The computation is done based on certain underlying assumptions of the scoring procedure. After administering the tryout, it is preferred to analyze the quality of test items since the quality of the test depends on the quality of each test item.

Commonly, item analysis provides information concerning how well each item in the test functioned. It also can tell us if a norm-referenced item was too easy or too hard how well it discriminated between high and low scorers on the test, and whether all of the alternatives functioned as intended (Grounlund, 1977:110). It means that item analysis provides information related to item difficulty, item discrimination, and effectiveness of distractors.

Item difficulty refers to the proportion of test takers who answer an item correctly (Flucher and Davidson, 2007:102). If most test takers can answer the item correctly means that the item is very easy and if most test takers cannot answer the item correctly means that the items is very difficult. It is generally assumed that items should not be too easy or too difficult for the population for whom the test has been designed. Items with facility values around 0.5 are therefore considered

to be ideal, with an acceptable range being from around 0.3 to 0.7 (Henning, 1987 in Flucher and Davidson, 2007:102). In other words, if the facility value of an item is around 0.5, it means that the item is not too easy and not too difficult which refers as an ideal level of difficulty.

A test item also should be able to differentiate between students who really possess the knowledge in answering the correct item and who do not. Flucher and Davidson (2007:103) explains that the responses to individual items are capable of discriminating between higher ability and lower ability test takers. To compute item discrimination, point biserial correlation is used to compute the association between responses to any specific item (i.e. a 0 or a 1). Items with an r_{pbi} of 0.25 or greater are considered acceptable (Flucher and Davidson, 2007:103).

To make an item able to discriminate between higher level and lower level students, it needs distractors. Distractors will distract test takers to choose incorrect answers. For lower level students, those distractors will be quite troublesome since they are not really sure with the correct answer of the item. In contrast, for higher level students, those distractors will be meaningless since they surely know the exact answer of the item. Further, a good distractor is able to divert students' attention to choose the correct answer.

The primary goal of distractor efficiency analysis is to examine the degree to which the distractors are attracting students who do not know the correct answer (Brown, 1996:71). A good distracter will attract more students from the lower

group than the upper group (Gronlund, 1977:113). The ability to influence students to choose the distractor will influence the discrimination value to differentiate lower level students and higher level students. To do this for an item, the percentages of students who chose each option are analyzed (Brown, 1996:71).

In conjunction with posttest instrument administered by Language Center (PUSBA) IAIN Raden Intan Lampung, it is urged to analyze the quality of the posttest instrument since it is not a standardized test. This test is made for the purpose of measuring students' achievement after joining matriculation program. Therefore, this research tried to investigate the quality of test items of posttest for matriculation program in terms of item difficulty, item discrimination, and effectiveness of distractors. The computation of item difficulty, item discrimination, and effectiveness of distractors will be analyzed by using ANATES.

METHOD

This research was descriptive research which tried to describe quantitatively the quality of test items used in posttest of matriculation program administered by Language Center (PUSBA). This research tried to analyze quantitatively the quality of test items in terms of item difficulty, item discrimination, and effectiveness of distractors. Since the posttest instrument used in the end of matriculation program for a few late terms was not the same, then, the posttest instrument which was analyzed for its items difficulty, item discrimination, and

effectiveness of distractor was the recent posttest instrument used for the first session of the third term.

The posttest instrument consisted of 50 items which were divided into two parts. The first part concerned grammatical mastery and the second part concerned vocabulary mastery. The students' score was obtained from the latest posttest administered on July 2015 by Language Center (PUSBA). There were 150 students' answer sheets analyzed to determine facility value, item discrimination index and distractor efficiency.

To compute item difficulty index (P), item discrimination index (D), and effectiveness of distractors, ANATES version 4 was employed for practicality and efficiency.

FINDINGS

The computation was conducted by using ANATES version 4. Although ANATES provided information related to reliability, lower level students and higher level students, item discrimination, difficulty level, correlation between score of the item and total score, effectiveness of distractors, this research only focused on item difficulty, item discrimination, and effectiveness of distractors.

Item difficulty

Item difficulty can be found out by computing facility value. The facility value ranges from 1 – 0.0, the closer the facility value to 0.0, the more difficult the item. In computing facility value, ANATES was used for practicality and efficiency.

The result of computation using ANATES demonstrated that there was one item (2%) categorized as very difficult item. This item is not good to be tested to the students since the facility value was 0.113 which meant that this item was out of tolerable range for facility value. This item was needed to be revised if it would be used in measuring students' ability. Further, there was one item (2%) categorized as difficult item which its facility value was 0.193. This meant that the facility value of this item was out of tolerable range. This item was not good to measure students' ability. In this case. This item should be revised for further use. Next, there were twenty-three items (46%) categorized as moderate item. The facility value ranged from 0.346 to 0.686 which meant that those item are in tolerable range. Those items are good to be used in measuring students' ability appropriately. Then, there were seventeen items (34%) categorized as easy item. The facility value ranged from 0.713 to 0.833. It meant that those items were out of tolerable range. Those items were needed to be revised in order to be used for measuring students' ability. Finally, there were eight items (16%) categorized as very easy item. The facility value ranged from 0.873 to 0.973 which meant that those items were needed to be revised. Those items were not good in measuring students' ability.

Item Discrimination

The computation of item discrimination was also by the means of ANATES. Item discrimination tried to differentiate lower level students, students who do not have the knowledge of the answer of the questions, from the higher level students, students who have the knowledge of the answer of the questions. Item discrimination value could be seen from the result of r_{pbi}.

The result of computation using ANATES showed that there were two items (4%) which their discrimination values were minus (-), -0.17.50 and -0.750. Those two items were needed to be revised. If the discrimination value was negative, it meant that the item was false in discriminating lower level and higher level students. The students which had knowledge of the question would be regarded as the students who did not have knowledge on the question if the discrimination value resulted was minus (-). Next, there were twelve items (24%) which their discrimination values were below 0.250. It meant that those items should be revised since those items could not differentiate between lower level students and higher level students. Finally, there were thirty-six items (72%) which their discrimination values ranged from 0.250 to 0.825. It meant that those items were successfully able to discriminate lower level students and higher level students. Therefore, there were some items needed to be revised.

Effectiveness of Distractors

Still, the analysis of distractor efficiency was computed by using ANATES. This analysis tried to figure out whether or not each distractor worked well in attracting students who did not know the answer of the answer of the question.

The result of the computation illustrated that only six items (12%) which all distractors worked well. This meant that each distractor in those items could influence the students who did not know the answer to choose it. In contrast, there were ten items (20%) which all distractors did not work at all. All students were able to choose the correct answer, including the students who did not know the correct answer of the items. Therefore, those distractors of each item should be revised. Finally, there were thirty-four items (68%) which only some distractors worked. Only some distractors were chosen by students and some of them were not.

DISCUSSION

Referring to the result of computation by using ANATES to finding out item difficulty, item discrimination, and effectiveness of distractors, it could be seen that almost test items (46%) were in moderate range in term of item difficulty. Most items were not eligible to be given to the students in measuring their ability or performance. Further, in term of item discrimination, it was almost 72 % items could be able to discriminate students who possess the knowledge in answering the question of the items. The scores on the whole test are the best single estimate of ability for each student. In fact, these whole test scores must be more accurate than any single item because a relatively large number of observations, when taken together, will logically give a better measurement than any of the single observations (Brown, 1996:68). Therefore, the better the items discriminate students' ability, the more accurate the results demonstrate students' real ability.

Concerning effectiveness of distractors, ten items (20%) which their distractors did not work at all should be revised. Revising such distractors would improve the item discrimination ability to differentiate students' who are able to answer the question and who are not.

In this case, it seemed that the problems arose because the test was not constructed based on test specification. It was not found that there was test specification as a guideline in developing the test. Therefore, it was impossible to determine the validity of the test, content validity and construct validity.

CONCLUSION AND SUGGESTION

Referring to the aforementioned discussion, the quality of posttest instrument used to measure students' achievement after joining matriculation program at Language Center (PUSBA) needed to improve since the level of difficulty for most items was out of tolerable range. In term of item discrimination, most items were good to discriminate students who are able to answer the question and who are not. In term of effectiveness of distractors, this posttest instrument should also be revised since only twelve items which its distractors worked well.

To improve this, it is suggested that the test should be constructed based on test specification to make sure that the test covers all sample materials to be investigated. By providing test specification will grant content validity of the posttest instrument. Further, such item analysis also should be conducted to make

sure that the constructed test is reliable to be given to the students in different batches and for the betterment of the quality of posttest instrument.

REFERENCES

Alderson, J.C., Clapham C., and Wall, D. 1995. *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press

Brown, J.D. 1996. *Testing in Language Program*. Upper Saddle River: Prantice Hall Regents

Flucher, G. and Davidson, F. 2007. *Language Testing and Assessment: An Advanced Resource Book*. Oxon: Routledge

Gronlund, N.E. 1977. *Constructing Achievement Tests* (2nd Edition). Englewood Cliffs: Prentice Hall Inc.

Gronlund N.E. and Waugh, C.K. 2009. *Assessment of Student Achievement* (9th Edition). Upper Saddle River: Pearson Education Ltd.

Heaton, J.B. 1999. *Writing English Language Tests* (New Edition). London: Longman Group

Huges, A. 2003. *Testing for Language Teachers* (2nd Edition). Cambridge: Cambridge University Press

APPENDIX

The Result of Computation Using ANATES for IF, ID, and Distractor Efficiency

Item Number	IF %	ID %	Options			
			a.	b.	c.	d.
1.	19.33	-17.50	+	**	---	--
2.	38.00	45.00	**	-	-	++
3.	79.33	35.00	-	**	++	_
4.	73.33	10.00	-	--	+	**
5.	34.67	52.50	++	**	-	-
6.	39.33	25.00	++	-	+	**
7.	76.00	42.50	++	++	**	+
8.	44.67	12.50	+	**	--	+
9.	53.33	72.50	++	++	++	**
10.	54.67	70.00	**	--	-	++
11.	42.67	77.50	--	--	**	++
12.	62.67	47.50	+	+	--	**
13.	80.00	52.50	**	++	-	--
14.	52.67	10.00	--	**	--	---
15.	78.67	40.00	**	++	++	+
16.	62.67	27.50	+	**	++	-
17.	42.67	25.00	-	**	---	--
18.	35.33	12.50	--	---	+	**
19.	51.33	0.00	--	**	+	-
20.	73.33	32.50	**	++	-	--
21.	82.00	-7.50	**	---	--	+
22.	71.33	32.50	-	---	+	**
23.	68.67	25.00	-	++	-	**
24.	35.33	47.50	-	**	+	++
25.	52.67	50.00	+	**	-	--
26.	91.33	32.50	**	--	--	---
27.	74.00	72.50	--	**	---	--
28.	11.33	5.00	--	-	…	**
29.	82.67	50.00	--	-	---	**
30.	48.67	30.00	**	++	-	+
31.	80.00	42.50	-	++	**	-
32.	87.33	5.00	-	**	++	--
33.	96.67	5.00	++	**	++	+
34.	67.33	82.50	-	-	**	++
35.	60.00	32.50	--	---	-	**
36.	67.33	37.50	--	**	---	-
37.	72.00	45.00	+	---	**	-
38.	72.67	70.00	**	---	-	--

39.	72.00	70.00	+	++	+	**
40.	56.67	72.50	--	**	--	++
41.	82.67	57.50	+	+	--	**
42.	94.67	15.00	-	+	**	+
43.	80.67	45.00	--	**	++	-
44.	70.67	42.50	**	-	-	---
45.	89.33	25.00	**	++	++	+
46.	83.33	10.00	--	--	++	**
47.	61.33	27.50	---	**	++	--
48.	96.00	10.00	-	**	+	++
49.	64.67	25.00	---	--	**	--
50.	97.33	7.50	--	+	---	**

Note:

** : Answer Key
++ : Very Good
+ : Good
- : Not Bad
-- : Bad
--- : Very bad