LLT, e-ISSN 2579-9533, p-ISSN 1410-7201, Vol. 20, No. 2, October 2017

Language and Language Teaching Journal
http://e-journal.usd.ac.id/index.php/LLT
Sanata Dharma University, Yogyakarta, Indonesia

TEACHERS' BEHALF ON LANGUAGE TEST CONSTRUCTION

Nyak Mutia Ismail and Marisa Yoestara
Syiah Kuala University and Serambi Mekkah University, Aceh, Indonesia
nyakmutiaismail2010@gmail.com
DOI: doi.org/10.24071/llt.2017.200201
received 12 May 2017; revised 15 July 2017; accepted 10 August 2017

Abstract
There are four steps necessarily to be conducted when designing multiple-choice test items, namely setting the objective, building both concise stems and options, determining one correct answer, employing item indices to accept or discarding items (Brown, 2004). As a matter of fact, most teachers in Aceh are not very well-informed about the fourth step and they accept all items as they are. This study focuses on high school teachers who undergo all of the steps offered in the framework when constructing multiple-choice items for English summative test(s). The qualitative method using framework analysis was used in obtaining the data. A questionnaire was distributed to 15 teachers. The analysis process was carried out through three-step analysis (Miles, Huberman, & Saldana, 2013). The results depict that the teachers hardly conduct the index determining step or try-outs when constructing a test. This implies that there is no empirical warrant that all items are worth tested and can be the fundamentals for decision-making when assessing and evaluating students' test results.

Keywords: test construction, language testing, multiple-choice items, summative test, and assessment and evaluation.

Introduction
Tests have long been used in any scholastic disciplines to measure learners' ability in their cognitive achievement, including in language teaching. Tests inform language teachers and instructors how far their students have mastered the materials taught prior to the test(s). Then the test result is used as an assessment parameter in setting their students level of ability—whether they are high- or low-performance students. Further, these assessments are used as hallmarks for wider domain in teaching and learning process which is evaluation. Evaluation could affect not only impersonal teaching policy such as technique and methodology applications, classroom managements, and teaching-material selections; but also imprint on the refinement of curriculum—the holistically massive setting of pedagogical policy. Thereunto, teachers and instructors are demanded to construct qualified test to administer to their students.

Particularly in language testing, English teachers nowadays design more multiple choice items in school summative test. Some considerations are worth to be take into account. First, these items are the ones consuming much time in designing and even can promote guessing and cheating during the administration process (Hughes, 2003). Second, from a survey done by the authors through

several schools in Banda Aceh with the English teachers, scarcely do teachers at school carry out a test try-out for a summative test, none of them were known to revise/opt out the unqualified items. Some schools even reuse the items for more than three years without knowing whether the items are indeed qualified for the students.

From some previous studies done in Aceh about multiple choice test item construction, below are presented three of them. The first one is a study by Setiyana (2016) at MAN Boarding School, Meulaboh. Her study focused on finding the validity, reliability, and item analysis of the test items. She employed checklists and document analysis during her data collection. The result showed that the test validity was poor but the reliability was high. Meanwhile the index difficulty was mostly easy; the discriminative index was also good; and more than 50% of the distractors were effective. The second one is a study by Khairunnisak (2016) which was investigating the validity and reliability of summative test in SMAN 1 Gandapura, Bireuen, Aceh, Indonesia. She employed a content analysis which worked fully on examining the multiple choice test item designed by the teachers at the school. The findings suggest that the items which tested reading comprehension were valid, but the items testing writing skills were not valid. Besides, the item indices from the items were also scrutinized. In item difficulty, she found that most items are at the easy level, but the discriminative index and distractors' efficiency were sufficient. Lastly, a study conducted by Syahputri & Ismail (2017) about construct validity of summative test items in a high school in Nagan Raya, Aceh, Indonesia. This study aimed at finding out whether the summative test items were compatible with the curriculum and syllabus offered by the Indonesia ministry of education. A qualitative design was employed; and in data collection process, the authors collected data through both analyzing the summative test content and interviewing the English teachers who designed the test at the school. From the findings, it was figured out that the summative test items in a Nagan Rayan high school were compatible with the national curriculum. From the interview, it was informed that the teachers also do process evaluation, instead of test evaluation alone, as suggested in the Indonesian curriculum 2013.

In further attempt to provide empirical data on this case, the authors are earnestly shedding light to the following question: In constructing multiple choice items, what steps are frequently skipped by high school English teachers in Banda Aceh?

Good Test Criteria

The characteristics of good items should inquire validity, reliability, as well as test piloting and revising where item analysis is satisfied (Qu & Zhang, 2013). Undeniably, the guide-setting process in a test construction is definitely critical since the test validity and reliability are intended as the core qualification for a test to be feasible (Haladyna, 2004; Cunnigham, et.al, 2013). Test validity is basic to any kinds of test-items as it really measures what the test is designed to measure, not any issues out of it (Cyril, 2005). In general, there are two points that most Acehnese teachers have seen as test validity, which are face validity and construct validity. The initial is a certainty that students do have knowledge on the test

items. And the latter is that the items are on the curricular syllabus—but Nunally (1972) refers to this as a case of reliability. Indeed, this is only the beginning steps of determining the test validity as the per-item validity has to be determined again. This is what most Acehnese teachers are lacking information about.

Next, it is test reliability which shows that the students' score remain particularly in a certain score-spectrum without drastic score-loss or gain. For example, a student scores 77 in a test, and two weeks later she/he scores 80. This score increase is still considered reliable unless there is an extreme increase. The test reliability is an index on where the final decision is based and this is a prerequisite to validity. A teacher cannot base his/her decision on a test which is not both valid and reliable. Hughes (2003) urges two conditions that can collapse test reliability, they are 1). The interaction between the examinees and the test since human beings are not machines and there is a minimum possibility a person can score in the same score-range after several time span; and 2). It is the scoring system—especially for essay items—which solely involves human beings, too, in the process. Regardless of these conditions, an unreliable test is hardly worth anything (Chiedu & Omenogor, 2014).

Steps in Test Construction

Cohen & Wollack (2015) suggest three general steps in a test construction. The initial step is *preparing the blueprints*—the process in which the purposes and objectives of the test are determined. This step is crucial since lack of blueprint preparation may lead to opaque test objectives. Next, it is *Item Development*. In this step, several test items are designed corresponding to each test objective(s). The latest to this is defining *Item Format* in where the test items will be intensified in multiple-choice, essay, cloze-test, or other formats. In addition, balance of all test objectives should be really deliberated in the last process.

In regard of particularly detailed steps in constructing multiple-choice items, Brown (2004) offers four steps. The foremost step to deal with designing the specific objective of the test items. For instance, when the specified topic is about grammar, the test designers should narrow the subtopic whether to test the *yes/no or wh-questions*, *word orders*, *direct/indirect speech*, *passive voice*, and so forth. Secondly, it is essential to design both stems and distractors in the simplest and most direct way. In a multiple choice question, a stem is the first introductory part where the examinees look for the intended answer, and distractors are wrong options excluding one and only single correct answer. The stem is not necessarily to be so long and intricate that the examinees might improvise—or even lose—their concentration on understanding the stem rather on deciding the correct answer while doing the item(s). The third move is to ascertain that there is only one exclusively correct answer, without any other possible correct answer instead. The last step is to try-out the test items to see their item indices. From the indices, test designers can decide whether to approve, revise, or override the item(s). There are three elements in item indexing, *i.e.:* item facility—to inspect the items' difficulty, item discrimination—to examine the items' ability in discriminating higher-group and lower-group students, and distractor efficiency—which shows effectual distractors that tempt the test-takers, especially from the lower group.

For a distractor to be claimed efficient, it has to be chosen by at least 2% of the whole test-takers (Fulcher & Davidson, 2007).

Method

The research methodology employed was basically qualitative design where the authors searched for the data quality instead of quantity. To be more specific, the framework analysis was used during the data collection. Based on the framework for multiple choice designed by Brown (2004), Table 1 provides some questions distributed to high school teachers during the data gathering. The premises were in Bahasa Indonesia and the respondents should answer *yes* or *no*.

Table 1. Questionnaire premises (developed based on Brown (2004))

1st principle	I determined the test purpose (remedial, formative, summative, etc)
	I determined the test objective (speaking, writing, grammar, etc)
2nd principle	I wrote the stems by directly citing from textbooks.
	I wrote the stems in direct and simple sentences.
	I wrote the distractors by directly citing from textbooks.
	I wrote the distractors in direct and simple sentences.
	I wrote the distractors in approximately similar length.
	I wrote the distractors in homogeneous part of speech.
3rd principle	I only designed one single correct answer without any possibilities for ambiguity.
4th principle	I tried out the test items I have designed.
	I applied item analysis (index of difficulty, index of discrimination, and distractor's effectiveness)
	I revised or discard the items with poor index.
	I decide type of scoring I would implement.
	I give feedback to students after the test.

The questionnaires were distributed to 15 high school teachers in Banda Aceh. Nine of them were senior high school teachers and the rest was junior high school teachers. They were chosen as the respondents of this study because they had been teaching high school for more than three years and they had designed various tests as well, including formative, summative, or remedial tests in the form of multiple-choice, essays, cloze-test, and so on. The data collection process was carried out within March-May, 2017.

Findings and Discussion

This section narrates the findings on this study followed by some theoretical grounding in the discussions. Figure 1 shows the findings of the steps that are done and not done by the teachers when constructing English test. Q stands for question, referring the ones to the questionnaire, blue bar represents the answer *yes*, and the red bar represents the answer *no*.

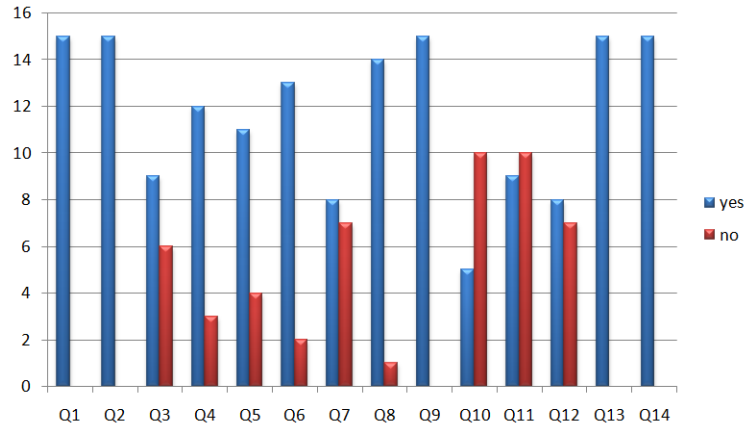

Figure 1. Steps done by teachers in constructing English tests

Figure 1 shows the total number of participants who determined the purpose and objective of test before they design it. Determining the purpose and objective of a test is a part of the first principle. It can be seen that all teachers perform step one and two, namely identifying the purpose and objective of the test. As mentioned by Jabbarifar (2009), setting purposes and objectives before designing a test is important because it lines out the rationale why a test is constructed, how a test is going to be administered, and what activities are going to be carried out in a test.

Next, the bars showing the responses or question three to eight reveals that there are some teachers who do not follow the rules offered in the second principle in language test construction. In response to Q3 and Q5, six teachers cited directly the sentences from textbooks and four of them also cited the distractors directly. According to Brown (2004), it is not suggested to quote both stems and distractors directly from textbooks without modifying them. Then, in Q4, three teachers do not design stems in direct and simple sentences. The response is similar to Q6 where two teachers do not make the sentences clear and simple in the distractors. Answering Q7, seven teachers informed that they do not write the distractors in similar length. In Q8, only one of them who does not write the distractors in a homogenous part of speech. Burton et al. (1991) urge that the sentences used in the stems and distractors of multiple choice items should be clear and concise. The sentences do not have to be complete. The following is an example taken from Burton et al. (1991, p.10).

A market clearing price is a price at which:
a. Demand exceeds supply.
*b. *Supply equals demand.*
c. Supply exceeds demand.

The example shows that the stem is not in a complete sentence, and neither are the distractors. The sentences are also in similar length and have the same part of speech – a simple sentence pattern of Subject-Verb-Object is employed in the example. This is considered as the directness and precision of the multiple choice

test items. In an objective test like multiple choice, the stems should be clear and concise (Zimmaro, 2010). Later, the response to Q9 shows that all teachers only design one and only best answer in the test. This is a framework offered in the third principle. Brown (2004) and Zimmaro (2010) also add that in multiple-choice, there should only be one correct answer while other options only act as the distractors.

Finally, in the fourth principle, Figure 1 shows the most striking red bars in Q10 and Q11. There are 10 teachers who do not undergo these steps, namely trying-out the test items and determining the item analysis. In Q12, we can see 7 teachers do not revise their test items. Since there were three teachers (from 15 respondents) who work at a private school, the authors were informed that in their school they have trainings and advisory boards for testing development. Therefore, they have to do the try-outs, determine the item analysis indices, and revise the test items before administrating the tests. On the contrary, there is no such information from the other teachers who work at public schools. Apparently, more ventures should be done to increase teachers' competence in Indonesia—including test development competence. As it is surmised by Rahman et al. (2015), based on teachers' competence test result in 2014, Indonesian teachers still face serious challenges in content knowledge. When they are still struggling with the content knowledge, it is very unlikely for them to succeed in developing test items, especially in trying-outs and revising. However, the teachers informed that they do decide the types of scoring systems they use beforehand and give feedback to students after the tests which can be seen in response to Q13 and Q14.

Conclusion

Since this study investigates which principle(s) Acehnese teachers mostly ignore during the test construction, there are two conclusions that can be drawn from the results. First, most teachers in Aceh are not accustomed to performing the fourth principle given by Brown (2004) in designing multiple choice test items, namely examining the item analysis. Second, the teachers do not try-out the test items because they are not well-informed about doing the try-outs with the test items and examining the validity, reliability, and other indices of the test items.

These results imply that most teachers only copy and reuse the test items from year to year without knowing whether the items still fit the current students' ability. A mild suggestion might be addressed to high school stakeholders to organize trainings on language testing and evaluation for their teachers.

Finally, future studies can ponder on steps that teachers use when designing other types of test items, such as essays, cloze-tests, matching, true-false, and so on. Besides, the verification on how the teachers conducted the subjective scoring with such tests may also be interesting to study.

References

Brown, H. D. (2004). *Language assessment: Principles and classroom practices.* White Plains, NY: Longman.

Burton, S. J., Sudweeks, R. R., Merrill, P. F. & Wood, B. (1991). *How to prepare better multiple-choice test items: Guidelines for university faculty.* Utah, Brigham Young University Testing Service.

Chiedu, R. E. & Omenogor, H. D. (2014). The concept of reliability in language testing: issues and solutions. *Journal of Resourcefulness and Distinction, 8* (1).

Cohen, A. S. & Wollack, J. A. (2015). *Handbook on test development: Helpful tips for creating reliable and valid classroom tests.* Retrieved from: https://testing.wisc.edu

Cuningham, S. A., Callaham, S. M. & Feld, J. K. (2013). *Item development and test construction guidelines.* Tucson, AZ: Assessment Technology, Inc.

Cyril, W. J. (2005). *Language testings and validations.* Hampshire, UK: McMillan Publishers.

Fulcher, G. & F. Davidson. (2007). *Language testing and assessment*. Boston: Routledge.

Haladyna, T. M. (2004). *Developing and validating multiple-choice items.* Mahwah, N.J.: Lawrence Erlbaum Associates.

Hughes, A. (2003). *Testing for language teachers.* Cambridge, UK: Cambridge University Press.

Jabbarifar, T. (2009). The importance of classroom assessment and Evaluation in educational system. In *The Proceedings of the 2nd International Conference of Teaching and Learning, INTI University College, Negeri Sembilan, Malaysia.*

Khairunnisak. (2016). *The item analysis of English summative tests* (Unpublished master thesis). Syiah Kuala University.

Nunally, J. C. (1982). Reliability of measurement. *Encyclopedia of Educational Research, 4*, 15-16.

Qu, W. & C. Zhang. (2013). The analysis of summative assessment and formative assessment and their roles in college English assessment systems. *Journal of Language Teaching and Research, 4*(2), 78-92.

Rahman, B., Abdurrahman, A., Kadaryanto, B. & Rusminto, N. E. (2015). Teacher-based scaffolding as a teacher professional development program in Indonesia. *Australian Journal of Teacher Education, 40* (11).

Setiyana, R. (2016). Analysis of summative tests for English. *English Education Journal, 7*(4), 433-447.

Syahputri, V. N. & Ismail, N. M. (2017). English summative test and national curriculum: The compatibility. In *The Proceeding of the 6th Aceh Development International Conference, Kuala Lumpur, Malaysia.*

Zimmaro, D. M. (2010). *Writing good multiple-choice exams.* Retrieved from: https://facultyinnovate.utexas.edu/sites/default/files/documents/Writing-Good-Multiple-Choice-Exams-04-28-10.pdf