Designing a more communicative grammar test

Ella Wulandari, M.A.
Yogyakarta State University
wulandari.ella@uny.ac.id

Abstract

In test design, practicality and validity have often been a dilemma. This paper presents the processes of designing and trialling out a test of grammar which was designed to be high in construct and content validity, innovativeness (communicativeness and interactiveness), potential to create positive backwash, and to some extent, practicality. From the result of the test trial, the test was then criticized for its Facility Value (FV) and Discrimination Index (DI), based on which some suggestions of improving the test design are offered.

Urgency of the test design

Many studies have investigated the prevalent use of multiple-choice (MC) tests and their effects in English language teaching (ELT), in English as Foreign Language (EFL) contexts (e.g. Li, 1998; Gorsuch, 2000; Yashima & Yoshizawa, 2006; Liu, 2007). In such contexts, testing grammar seems indispensable to teaching and learning and has been closely associated with MC technique (Dávid, 2007). Dávid further states that MC-based grammar testing is less appropriate since the concept of grammar has been redefined to involve not merely syntactical but also semantic and pragmatic elements. So are they for assessing grammatical ability, defined as the integration of grammatical knowledge and strategic competence, which manifests in capability to utilize grammatical knowledge accurately and meaningfully in testing or real situations (Purpura, 2004, p.86).

However, MC items still play a major role in grammar testing in EFL contexts, including in Indonesia, for its economy, practicality and reliability ((Dávid, 2007), despite their being prone to test method effect. A recent study on language structure tests by Currie (2010) finds that MC formats are

> "likely to elicit a greater proportion of format-related 'noise' than the language performance actually sought, and that the results of such tests are unlikely to fully reflect the responses that the test takers would offer if the test were set in a constructed-response format. Arguments concerning the objectivity of the multiple-choice format, and its

> practicality as a method of testing large numbers of learners efficiently and cheaply must therefore be weighed against the risk that the measurement of the intended construct is likely to be contaminated by the effect of the item format" (p.487).

Given this, Currie implies the need to employ constructed-response format to reduce 'format-related noise' or test method effect potentially resulted from MC items, as 'it encourages guessing' (Heaton, 1988, p.26). Such items also often give misleading indicators of one's language proficiency since they merely assess grammatical knowledge and do not require using language in communicative functions (Purpura, 2004). The less communicative grammar tests, as commonly found in Indonesia, is the issue with which the test design discussed in this paper is mostly concerned. The designed test (see Appendix 2), summative test for "Structure II" class, is therefore an attempt to provide a different type of grammar testing in an Indonesian university. The test, where greater emphases are given to its communicativeness, is to narrow the gap between the 'construct' of grammatical ability and MC-dominated grammar tests utilized in that particular educational setting.

Testing context

The context where the designed test will be used is one of "Structure II" classes in English Education Department, State University of Yogyakarta, Indonesia. The class has 20 students, who were in semester 3 back in 2010, and major in English Education. Currently, they are doing "Structure III", as the university is running odd semester. The department has two study programs, i.e. English Education and English Literature, and institutionally aims at producing teachers and professionals in English (education). Grammar is taught separately under "Structure". To date, grammar assessment in the department has been misunderstood as simply assessing learners' knowledge of linguistic forms in isolated sentences or contexts, despite its contradiction to the course objectives. Explicit grammar teaching takes place in "Structure" class, going from level I to IV, indicating stages of learning areas and semester when it is offered.

To go to higher level of "Structure", students have to pass the preceding class, determined by their result of achievement tests, including formative and summative tests. The tests, to the writer's own experiences as one of the teaching staffs at the department, are mostly MC and/or discrete items. Those items are well known among the staffs, as they are easy to

score, or practical. Though MC items could extremely be difficult to design (Dávid, 2007), the teachers 'tackle' this difficulty by adopting from test banks available online or in textbooks, without any validation measures. Unchecked for its validity, grammar tests in "Structure" appear to lack construct validity. It insufficiently assesses students' capacity to use their grammatical knowledge in test-taking or real situations, and therefore does not precisely indicate their grammatical ability. Though low-stake, less valid summative tests might result in passing or failing wrong students, and tend to bring negative backwash to teaching.

Test content

As a classroom test, the test content can be based on the course's syllabus (Appendix 1), or syllabus-content approach test, or course objectives (Hughes, 2003). To base the test content on the syllabus seems to make the test fair to the students and therefore high in face validity, as what is examined is what it is taught in the class. However, Hughes argues that it is advisable to rely the test content on the course objectives or aims, since it will present precise information about achievement of individual and whole class, and make positive backwash on teaching possible. In particular, he suggests that it will save students from a badly designed syllabus or poor teaching practice, which does not correspond with the course objectives. In this regards, the test writer uses the syllabus content as the basis of test content, but takes the course objectives into account. The syllabus states that "Structure II" aims at 'developing students' knowledge and ability to use intermediate structure of English at both receptive and productive levels' (See Appendix 1). As a result, the test is designed to be able to assess students' receptive and productive grammar competence, or in other words, grammatical ability.

There are nine sections with a number of items in the test. It attempts to include most the syllabus content that is form-based, and is designed to take 90 minutes. The arrangement of the test is as follows:

Sect	Focus on/ expected use of form	Question/ Instruction	Use of text(s)	No. of items	Weigh/ mark per item
1	Simple past, past perfect, past perfect continuous, and past future tenses.	Simple past, past perfect, past perfect continuous and past future tenses. *Below is a story that happened in the past. Circle the letter to choose the appropriate form of the verb. The first sentence has been done for you.*	Story (from textbook)	10	1x10 =10

2	Past simple or present perfect tense	*Jane, your Australian friend, is going to visit Yogyakarta. You want to inform her about the following recent news about Mt. Merapi. Complete the news by changing the verbs on the right side into* ***past simple or present perfect*** *tense.*	News, authentic adapted	10	1x10 =10
3	Present future tense (to be going to, will), main and auxiliary verbs, and noun phrase (infinitive, gerund).	*Jane is now visiting Yogyakarta for a week. She asks you to be her tour guide. Unfortunately the weather forecast of this week does not look good.* **a.** *First, write* ***three*** *sentences reporting the weather forecast of the week. You may or may not use the words in the table.* *Example: The clouds are going to go away tomorrow.* **b.** *Now, can you think of activities that will make her enjoy her stay in Yogyakarta? Considering the weather forecast of the week, write six sentences about* • *3 activities that are good to do during the week, and* • *3 activities that are not good to do during the week.* *Example:* *Watching movies on Saturday night is a good idea.* *It is a bad idea to go to beaches on Sunday.*	Weather forecast report authentic adapted	9	2x9=18
4	Modal (must, should, can, need), main and auxiliary verbs.	*To help Jane become familiar with Yogyakarta, describe some common road signs below with your own sentences using* ***must, should, can*** *and* ***need****. ONE sentence each.*	Pictures, authentic (2 from Jogja)	3	2x3=6
5	Modal (must, need)	*The following are some everyday situations in Yogyakarta that might also be unfamiliar to Jane. Describe the situations to her with your own sentences. ONE sentence each.*	(context-building) stimulus	2	2x2=4
6	Main and auxiliary verbs, modal (used to), and negative statements	*But the road signs and situations above <u>were</u> not common to Yogyakarta people 20 years ago. Using* ***used to, hardly, rarely, seldom*** *and* ***never****, describe how Yogyakarta has changed in some ways. Example: (picture texts)*	Pictures, authentic (2, from Jogja)	10	2x10=20

	(hardly, rarely, seldom, never)				
7	Inversion of negative statements)*	*Then change FOUR of your sentences on the right column using **inversion**. Write ONLY the inversion. Example: (0) **Rarely do people ride bicycles nowadays**.*	In relation to no. 6	4	1x4=4
8	Main and auxiliary (modal can, have to) verbs, and interrogative sentences	*Jane wants to go to Jogja Library Centre. To help her obtain sufficient information, use the information on the library brochure to create a list of Frequently Asked Questions (FAQ) and answers. One question and answer for each topic. Example:* *Q: What kinds of book can I borrow?* *A: You can borrow all kinds of books except dictionaries, encyclopedias, and clippings.*	Brochure (made), the library can be found in Jogja	8	2x8=16
9	A number of possible forms, future perfect (continuous) tense	*You know Jane loves watching movies. The following is the schedule of the movies that are now playing in Cinema 21 Ambarukmo Plaza, Yogyakarta.*	Movie schedule, authentic	5	2x5=10
Total				61	98+2 (bonus) =100

Table 1. Test content

Marking criteria

This particular test employs both objective and subjective scoring. Sections 1, 2, and 7 of the test require no judgment, as there is only right/wrong answer. Thus, the items in the sections are scored one each. The following sections, however, are marked subjectively by the test writer, on form and meaning bases. Accuracy in form earns one mark and appropriacy in meaning obtains one score. Where human judgment and decision are called, there would never be an objective marking since "rating always contains a significant degree of chance, associated with the rater and other factors" (McNamara 2000, p.37). Similarly, Hughes (2003) points out that subjective tests will not reach reliability as perfect as that of objective tests, though some attempts can be made to enhance test score reliability.

To make its score reliability improved, the test employs a carefully designed answer key with samples of correct answers **(**see Appendix 3**)** and has a number of tasks and items (Purpura,

2004). Concerning items other than MC and gap-filling, Hughes (2003, p.178) emphasizes the importance "to be clear about what each item is testing, and to award points for that only". He further argues that non-grammatical errors should not reduce the score. Nor should errors in aspects of grammar which a particular item is not meant to test. Yet, besides section 7 which only assesses form accuracy, distinguishing grammatical from non-grammatical errors seems less appropriate for this particular test, as it involves form and meaning criteria of correctness. Though score is given to the tested form only, if the sentences or responses are ungrammatical, it might affect their meaning appropriateness as well. In the end, it is much practical to rely on form and meaning scoring standard, with correctness of each is awarded one score.

The next attempts to increase test score reliability are having more raters to rate the test and reasonable freedom of response variation. Since the class teacher refuses for inter-rater marking, the test writer cannot establish inter-rater reliability. Yet, having more independent judges in scoring the test in preferable. In addition, Hughes (2003) recommends not giving too much freedom or many choices of responses that students can produce. He argues that "the more freedom that is given, the greater is likely to be the difference between the performance actually elicited" (p.45). The test, therefore, gives several linguistic labels or examples of expected responses in order to reduce the differences of responses as well as to lead students to produce the expected forms.

The significant features of the test and their rationale

This test is significant in terms of its construct and content validity, innovativeness (communicativeness and interactiveness), potential to create positive backwash, and to some extent, practicality. Bachman & Palmer (1996, p.21) describe construct validity as "the extent to which we can interpret a given test score as an indicator of the ability(ies), or construct(s), we want to measure". This grammar test uses construct of grammatical ability proposed by Purpura, as mentioned earlier. It does not merely assess grammatical knowledge (sections 1, 2, 7) but also encourage students to maximize their strategic competence in utilizing their grammatical knowledge accurately and meaningfully in completing test tasks or performing real communicative functions (sections 3, 4, 5, 6, 8, 9). Its form and meaning scoring bases also show that it focuses on both receptive knowledge of structure and capacity to produce meaningful language through uses of grammatical knowledge. The test is also high in content validity, as it is

highly relevant with what is taught. Its content is based on the course syllabus and objectives, as explained before. Each of the tasks is made to likely elicit responses with the expected forms or structures. To ensure its likeliness and its score reliability as previously elaborated, each has either linguistic labels or examples of desired answers.

With regards to innovativeness, this test seems to provide an alternative to grammar testing in the department. First, it clearly has different format from the usual grammar tests, as it tends to have more balance between the aspects of real language-use and those of test tasks, or in short, it is, to some degree, more communicative (Bachman & Palmer, 1996). Its degree of communicativeness is achieved through utilizing a range of types of grammar tasks, lying on a continuum , including selected-, limited- and extended-response task types (Bachman & Palmer, 1996, pp.53-54; Purpura, 2004, pp.126-145). Its first two sections utilize selected-response (MC) and limited-response (gap-filling) tasks. The following sections, however, employ constructed-response tasks.

The first task is used to measure recognition skills dealing with grammatical form and/or meaning. It is scored right/wrong. Though it is the least communicative grammar task, selected-response through MC items can serve as a lead-in to the test. It helps students gain confidence by allowing them to tackle the easier questions first. To enhance candidates' confidence and greater chance for success in this particular task, explicit 'linguistic labels' that explain what forms are expected from students' responses are provided. Besides, their present is to reduce the range of variation of responses, as suggested by Hughes (2003, p.46). Though considered less communicative, the stems, or "the initial part of each multiple-choice item" (Heaton, 1988, p.28) are presented in contexts. Heaton further highlights the importance of contexts in MC format, suggesting that

> "[d]econtextualised multiple-choice items can do considerable harm by conveying the impression that language can be learnt and used free of any context. Both linguistic context and situational context are essential in using language. Isolated sentences in a multiple-choice test simply add to the artificiality of the test situation and give rise to ambiguity and confusion" (p.28).

Given this, the MC items used in the test are therefore provided in contexts through a narrative text (story). In fact, all items in the test are in context. Even in question number seven (see Appendix 2)**,** the responses are produced in relation to the previous question which is context-bound.

The second task, limited-response, appraises more elements of grammatical knowledge. It also uses one criterion of correctness, as "ideally, gap-filling items should have just one correct response" (Hughes, 2003, p.174). Gap-filling is used because it is a degree higher in communicativeness than MC. In that way, the test could go gradually becoming more communicative in the end. Such gradually increasing degree of communicativeness is important in the setting where the test is planned to administer, as the students (and the teachers) in the department have long been accustomed to selected- and/or limited-response tasks in grammar testing. MC and gap-filling serve as a lead-in not only to the test itself but also to this 'new' types of grammar tasks in the test. If the test does not involve any selected- or limited- response tasks, it is feared that the students cast doubts to the test's face validity, simply because it does not 'look' like a grammar test. Further, such doubts may lead to greater test anxiety among them which might affect their test performance. That negative response is surely undesired since the test is also seen as an introduction to the more communicative grammar testing.

This test is argued to be toward the communicativeness end of the continuum since constructed-response grammar tasks dominate the test. In such tasks, input are presented "in the form of a prompt instead of an item, involve language and/or non-language information", which is varied in quantity, and elicit responses which greater in length (Purpura, 2004, p.139). In particular, the test presents some prompts through different authentic text types including weather report, authentic pictures of road signs and Yogyakarta, the city discussed in the test, in the past, brochure, and movie schedule. The use of authentic texts is hoped to increase the test's communicativeness degree as well as to expose students to different language uses. Most importantly, the texts and (certainly) the tasks, are chosen for their interconnectedness in creating a 'background story' or 'theme' of the text suitable for the intended test-takers, which is assumed to make the test more interactive.

The test seems to be highly interactive for some reasons. First, it allows students to engage their cognitive and metacognitive knowledge (strategic competence) in making use of their grammatical knowledge accurately and meaningfully (Purpura, 2004, p.153), in completing the extended-response tasks in the test. Secondly, it enables candidates to draw on their topical knowledge about everyday life in Yogyakarta, as questions number 2 – 9 are given in relation to situations or happenings in the city, where the candidates live. It also builds on test-takers' positive affective schemata, for its 'topical' familiarity to them. Yet, Purpura cautions that,

though engagement with topical knowledge and positive affective schemata can create test's interactiveness, it risks the test's validity as measuring other than grammatical ability, if the topical knowledge has little relevance with the grammatical knowledge being assessed. The topical knowledge drawn on the test, however, seems to be highly relevant with the grammar task characteristics, and appropriate to meaningfully yield the expected answers from the candidates. At the very least, the test is still interactive in that it builds on a 'theme' that relates one question to another. In that way, it gives candidates a sense of experiencing the language tasks themselves in real-life situations. Further, the tasks (reporting, giving ideas or opinions, describing) are assumed to be 'possible' happening in an EFL setting, making them more meaningful.

Meaningful test tasks and authentic texts used in the test are crucial in making the test less artificial sample of language use. Testing in contexts will potentially create positive effect to teaching and learning, or backwash. Through the test, candidates are exposed to authentic language uses and contexts and thus able to see the relationship between the form(s) that they have learned, and how they could use them in real communication. The teaching will therefore not be suggesting artificial uses of language anymore, as usually found in teaching grammar in isolated contexts. Grammar teaching will then be more well-balanced between emphasizing grammatical knowledge and capacity to use the knowledge accurately and meaningfully in authentic discourse. In the long term, grammar testing and teaching culture in the department will not segregate grammatical competence from grammatical performance in meaningful contexts. In this way, the positive backwash is achieved, as the test meets the construct of grammatical ability, which it intends to examine. In short, it has big potential to bring about positive backwash to teaching and learning, as it has high construct validity.

In contrast to its high validity, the test's practicality is relatively low. Its practicality is, however, considered to be one of significant features of the test, as it is assumed that its level of practicality can easily increase after this type of test has gained acceptance and become part of grammar teaching and testing in the department. As stated beforehand, designing MC items is actually far more difficult. Besides, it measures only grammatical knowledge. Creating extended-response grammar tasks can be of little difficulty, once test-writers (or teachers) know exactly what aspects of grammar are going to be measured and have determined ways of testing them, certainly, in contexts. Practicality in scoring will be faster, as the test-writers become accustomed

with these types of tasks and have had ideas of what expected responses will likely be like. Surely, detailed scoring criteria with examples of correct responses are of great importance. In the department particular setting, the ease in scoring will soon be overcome, as the department has been establishing teacher group work, which is a cluster of teachers teaching the same subjects. These fellow subject teachers are potential inter-raters in marking processes.

Trialing, evaluation and test revision

The designed test was tried out in the fore-planned setting. There were, however, less number of students, 14. The test was administered by the class teacher and took 90 minutes, sharp. There was no significant technical or test administration difficulties found during the test. Yet, apparently, the pictures used in the test turned to be so poorly black-and-white copied that many candidates reportedly, frequently asked the teacher for clarification. They, later, expressed their complaint in the questionnaire attached to the test and completed right after they finished the test. The questionnaire is the test-writer's own idea, expecting that it might help her better understand how test-takers perceive the test, which is important in evaluating and improving the test.

Evaluated, the test has some drawbacks in terms of reliability (layout, instruction/task, and item ambiguity), and degree of difficulty (in relation to time and item weigh). Hughes (2003, p.47) suggests that the test should be "well laid out and perfectly legible", to increase its reliability. In the test, problem with unclear pictures is quite severe. One of candidates misinterpreted picture *a* in question 4 (see Appendix 2), and therefore provided a less appropriate response. Other pictures, though unclear, did not cause problems as big. The test overall is readable, but more spaces are needed in question 6. To improve its layout, the pictures in the test were enlarged or even changed. Pictures and texts, showing weather forecast for few days in a week, placed the weather forecast table. The first two pictures in question 4 were enlarged and the last two were substituted with new bigger pictures of public signs originally found in Yogyakarta. More spaces were added to write answers of question 6.

The next reliability issue of the test is instruction or task ambiguity. Hughes (2003, p.47) advises test-writers to "provide clear and explicit instructions". Question 4, 5, 6 and 9 in the test, however, might be confusing or unclear, as many candidates answered them less accurately or appropriately. The instruction of question 4 says,

> *"To help Jane become familiar with Yogyakarta, describe some common road signs below with your own sentences using **must, should, can** and **need**. ONE sentence each. (6 marks)" (see Appendix 2).*

Though the modal verbs given in the instruction can all well describe the provided road signs pictures, the use *'and'* instead of *'or'* made one candidate misinterpreted as describing each of the pictures using all the four modal verbs, even when *'ONE sentence each'* is explicated. This particular task is not meant to be too difficult, as shown by its score mean, 4.321 out of the total score, 6 (see Appendix 4).

The word *'describe'* in the instruction of question 5 clearly causes confusion among students. The task seems unclear as well, though it is not known whether it is the students who do not get the 'task' well. Very few candidates can answer the question accurately and appropriately. The mean of candidates' scores for this particular test is very low, 1.574, from the total score, 4. Responding to *'describe',* most candidates gave description from other points of view regardless the given prompts, suggesting uses of modal verbs to express certainty and necessity. Question 5 (see Appendix 2) says,

> *The following are some everyday situations in Yogyakarta that might also be unfamiliar to Jane. Describe the situations to her with your own sentences. ONE sentence each. (4 marks)*
> a. A man or woman in orange shirt or vest is helping you with parking in commercial public places like mall, stores or markets. They give you tickets. *(certainty)*
> b. Passengers in a car are not wearing seat belts. *(necessity)*

Further, the test-takers tricked the instruction '*ONE sentence each'* by writing a long single sentence, comprising at least one independent and one dependent clause, to give broad description of the situations prompted in question 5. To deal with this, the instruction and the task were modified into,

> *Jane also asks you about the following situations that she saw on the street. Using **must** and **need,** describe each of the situations in **a less than 10 words sentence**. (4 marks)*
> a. A man in orange shirt or vest was helping a girl with parking. He gave her a ticket. *(Who is the man?)*
> b. In a car, passengers on the back seats are not wearing seatbelt. *(Why?)*

Linguistic labeling is added to the instruction and an exact number of words substitutes *'ONE sentence each',* to elicit the expected responses, involving uses of modal verbs to express certainty and necessity.

Similar to question 4, the instruction in question 6 in the trialed test is unclear. The question expects candidates to produce sentences using *used to, hardly, rarely, seldom and never*. It actually intends to make examinees write 5 sentences using *used to*, to describe how Yoygakarta was like 20 years ago. It is also meant to measure candidates' ability to use *hardly, rarely, seldom and never*, to picture how Yogyakarta has changed nowadays, as shown by the example (see Appendix 2). Some students answered as expected. Some others, however, thought that they should also use *used to* to describe how the city has progressed recently, which is actually less appropriate, given its function to state habits in the past. Consequently, the instruction was modified into.

> *But the road signs and situations above <u>were</u> not common to Yogyakarta people 20 years ago. Using* ***used to,*** *describe Yoyakarta 20 years ago. Using* ***hardly, rarely, seldom, barely*** *and* ***never****, describe how Yogyakarta has changed nowadays* (see Appendix 5).

Likewise, the instruction of question 9 is confusing, as the word *'describe'* seems, again, distracting candidates from the intended task. This particular question is indeed feared for not measuring grammatical knowledge, but reading comprehension. However, the prompts, are presented in dot points, to ease comprehension. Still, only few students can come up with the expected or slightly appropriate responses, as suggested by its score mean, 4.5, from the total correct score, 10. This leads the test-writer to assume that the students found it hard because they had to complete too many questions within the given time, which will be discussed later.

Item ambiguity is found in the first two questions. Checked for its Facility Value (FV), or its difficulty level, and Discrimination Index (DI) (Heaton, 1988), most items in the questions are, unfortunately, low in both, or even unsuccessful to discriminate stronger from weaker students (see Appendix 4). Heaton (1988, p. 179) argues that for achievement tests, FV of 0.5 is advisable, though a range of 0.4 to 0.7 has been commonly accepted. In the test, only item number 6 and 10 of question 1, and item number 3 fall nicely around 0.5 of FV. The other items in question 1 has FV higher than 0.64, which means that they are too easy. Item number 4 has DI of 0.071, which indicates its high difficulty. Items number 1, 4, and 5 of question 2 have FV of 0.857, which is also too easy. Items number 9 and 10 even have FV of 1, which is undesirable in any tests. The other items have FV less than 0.3, which is extremely difficult. In short, FV of these two questions are less acceptable. Similarly, DI of the two questions is far from perfect. Sadly, only item number 1 of question 1 closely reaches acceptable DI of 0.45 (Heaton, 1988,

p.180), though its DI is only 0.43. The other items in questions 1 and 2 have DI much lower than the expected one, or even have zero DI and minus DI. Items number 4, 7, 8 in question 1, and item number 4 in question 2 have minus DI, which shows they discriminate the wrong way (Heaton, 1988, p.181). Moreover, items number 1, 5, 8, 9, 10 have DI of zero, or do not discriminate at all. In short, the FV and DI of the first two questions of the test show that the questions badly need revision.

Question 1, in particular, has a problem with the distractors of its MC-based items. Constructing good distracters that can really work to distract without putting candidates into a trap is difficult. Since the question uses linguistic labeling (simple past, past perfect, past perfect continuous and past future tenses), the distractors have to be choices of verb phrase with the mentioned forms. Thus, this makes the distracters narrow in variation. This later causes stronger candidates to think them as too easy, who therefore consider the distractors as a trap (Heaton, 1988, p.32). To resolve this problem quickly, the question is then changed into gap-filling. Meanwhile, question 2 is dropped, since it hardly discriminates the candidates (see Test Sheet). This may be resulted from the use of news text, written by non-native speaker (NNS) journalist, and still, adapted by the test-writer. Clearly, distinction between past simple and present perfect is very subtle for candidates to see and for the test-writer, who is a NNS, to understand. So, for a quick solution, the second question is eliminated. Further, reduction in number of items is required for improving the test's difficulty level, discussed next.

The other reason why the test needs improvement is its difficulty level. As stated before, the use of selected- and limited-response tasks is expected to well sequence its level of difficulty. Applying only form accuracy in the first two questions should also help proportionate its difficulty level. Yet, the test-writer thinks that the test might be too difficult because there are too many number of items in the given time. The subsequent questions except question 7 are increasingly difficult. The last question may be the most difficult, as it demands a bit of higher thinking, or analysis, when understanding the prompts given. They, however, all weigh two points, for form and meaning. Yet, the test-writer does not intend to give different weighs for those questions. To compensate with its increasing difficulty, she prefers to reduce one question, which is question number 2, for its poor (FV) and (DI), as discussed just previously.

Score interpretation

The scores of the candidates have been calculated for central tendency and standard deviation, and then tabulated (see Appendix 4). The mean (M) and median (Me) of the test are 60.929 and 61.5. The mean of the test is relatively low, in that it sits around the passing grade of the test, which is set as 60%. It implies that the students did not do quite well in the test. Concerning the median, Heaton (1988, p. 176) states that, "such a close correspondence is not always common and has occurred in this case because the scores tend to cluster symmetrically around a central point". Interestingly, the test has no mode, as each score is attained only by single candidate. To show the spread of the scores, the range of the test is calculated. Its range is 35.5, which is high. Yet, Heaton further argues that standard deviation (SD) is a more appropriate measurement for this. The SD of the test is **2.876,** which, according to Heaton, shows a smaller spread of scores. He further points out that

> "if the test is simply to determine which students have mastered a particular program of work or are capable of carrying out certain tasks in the target language, a standard deviation of 4.08 or any other denoting a fairly arrow spread will be quite satisfactory provided it is associated with a high average score" (p.178).

Given that the test, an achievement one, is to measure whether candidates have mastered a set of structures examined through different tasks, its SD is considered as acceptable.

Feedback received from the candidates

The test-writer apparently asked the candidates to complete a questionnaire written in Indonesian language, to better understand how they perceive the test. It is written in the language to avoid unnecessary confusion due to the use of English. It uses three-scaled of responses; 1 means 'very', 2 means 'rather/fairly', and 3 means 'not at all'. The questionnaire in both languages can be found in Appendix 5. The test-takers' responses as reflected in the questionnaire is concluded in the following statements:

1. Most candidates think that the test is neither difficult nor easy.
2. Most candidates consider the test's difficulty as well-graded.
3. Two candidates think that the test's instructions are very unclear, while the rest of them say they are fairly clear.
4. All students answer that they are either not accustomed to this type of test at all or rather familiar with it.

5. Around half of the candidates state that they prefer this test to the usual grammar tests they have had and that it is rather possible to actually use the tested forms in real communication, as suggested by the test tasks.
6. All of them agree that the test engages their topical knowledge and that the test is taken place in relatively sufficient time.
7. Most candidates regard that the test is well laid-out and appropriately assessed on form and meaning bases.

Given those statements, it is interesting to know that despite their generally low scores, the candidates do not think that the test is extremely difficult. Last, two concerns were given concerning pictures being unclear and insufficient time for completing the test.

In conclusion, regardless its weaknesses, the test seem to provide an alternative as a grammar test with a higher degree of communicativeness. Most importantly, its high construct and content validity is expected to create positive backwash to teaching and learning. It makes grammar teaching and testing more meaningful and therefore communicative. Still, more attempts and preparation in test construction are needed to improve the test. Using selected- or limited- response tasks in the test, though not widely recommended for communicative testing, should have involved careful considerations. Moreover, creating good distractors for MC format is shown not easy. Extended-response tasks are, therefore, advisable to assess grammatical ability, as used in the trialled (see Appendix 2) and revised tests (see Appendix 6).

REFERENCES

Bachman, L. and A. Palmer. (1996). *Language testing in practice*. Oxford: OUP.

Currie, M. (2010). The effect of the multiplechoice item format on the measurement of knowledge of language structure. *Language Testing*, *27*(4), 471-491.

David, G. (2007). Investigating the performance of alternative types of grammar items. Language Testing, *24*(1), 65–97. Doi.: 10.1177/0265532207071512

Heaton, J. B. (1988). *Writing English Language Tests.* London: Longman.

Hughes, A. (2003). *Testing for language teachers.* (2nd ed.). Cambridge: CUP.

McNamara, T. 2000. *Language Testing.* Oxford: OUP.

Purpura, J.E. (2004). *Assessing Grammar.* Cambridge:CUP

LIST OF APPENDICES (in the order they appear)