ISSN: 2252-8385 (p) 2541-4399 (o)
VISION

VISION: JOURNAL FOR LANGUAGE AND FOREIGN LANGUAGE LEARNING, 2019
VOL.8, NO.1,53-78
http://dx.doi.org/10.21580/vjv8i13313

Designing a Midterm Reading Test for Junior High School Students in Semarang: A Practical Experience of a Master Student in TESOL

Waliyadin
English instructor at the Language Development Center State Islamic University (UIN) Walisongo Semarang
waliyadin.nuridin@gmail.com

ABSTRACT

Designing a test is an uneasy task. It is proven by the fact that many teachers find difficulties in designing a good test. The present article attempts to review the English school exam in Indonesia and present an example of the test design of a midterm reading test for junior high school students in Semarang. To this end, firstly, a sample of test script of the school exam for JHS students in Semarang in the academic year 2017/2018 was reviewed. The review of the test shows that creating a good quality of test items still become a concern for English teachers as they made some grammatical errors. Consecutively, a midterm reading test was also designed and administered in a junior high school in Semarang, Central Java. Then, the test designed was evaluated by considering the validity, reliability, and practicality as suggested by (Brown & Abeywickrama, 2010). The administration of the test and evaluation show some weaknesses; therefore, the suggestions for the betterment of the test design are also made, such as revising some test items with a very low facility index and negative discrimination index. This study implicates pedagogically the teachers to train students on how to summarize and paraphrase since these skills underrepresented.

ARTICLE HISTORY
Received 17 February 2019
Accepted 30 April 2019

KEYWORDS
Testing and assessment; reading test; washback; school exam

Introduction

Testing and assessment have become an important issue in Indonesian education and attract attention for stakeholders to take action. In fact, as an attempt to improve the quality of education, the Indonesian Ministry of Education and Culture generates some educational policies regarding testing and assessment. Recently, through the Ministry of Education and Culture Decree No. 4/2018, a regulation concerning the testing and assessment conducted by schools and government was made. This decree changed the status of the school exam from low stakes to high stakes since the school exam and summative test in every semester are used as the basis to make decisions about the student graduation. Additionally, the results of the school exam according to the decree also can be used as one of the considerations for the selection for higher-level education as an added score of the national exam.

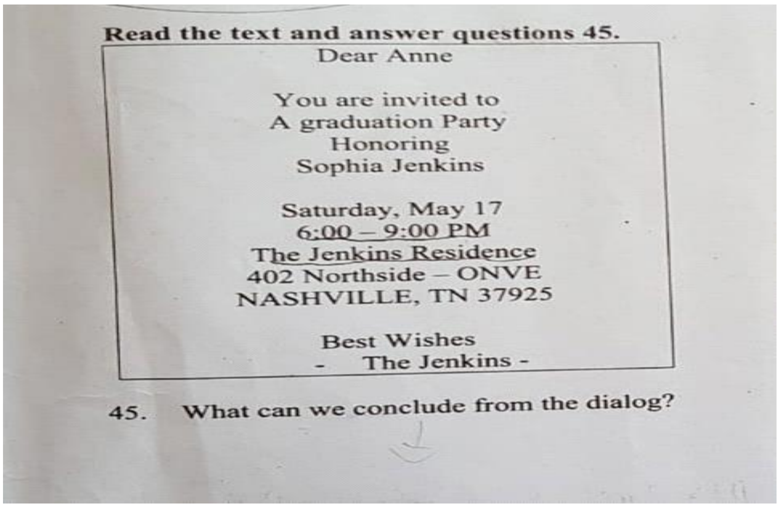
To begin with, it is important to understand the school exam in the Indonesian context. The school exam is categorized as an achievement test. According to the Ministry of Education and Culture Decree No. 4/2018, the school exam is called the national standard school exam which is defined as an activity to measure students' academic achievements and to acknowledge that students have achieved standard competence for graduation determined by the government. To this end, the Indonesian government through the board for the national standard of education (Badan Standar Nasional Pendidikan, *BSNP*) manages the quality of the test. In fact, the test specifications and about 20-25% of the test items are created by *BSNP* (see the Board for National Standard of Education Regulation No. 0045/BSNP/II/2018.

Literature Review

The school exam is conducted before the national exam; therefore, it has a function as complementary to the national exam because the national exam does not cover all aspects of student competencies. For example, the English curriculum mandates four language skills, including reading, speaking, listening and writing, but in fact, reading and listening are predominantly tested in the national exam while the other skills are ignored. Interestingly, the test items of school exam are constructed by school teachers in accordance with the specifications created by the board for the national standard of education and the indicators of test item are created by *MGMP* (a teacher association teaching in the same subjects).

The positive impacts of the school exam, particularly for English seems promising. This is because of combining the specifications of the test developed by the board for the national standard of education and the teachers' involvement in designing the test items. The inclusion of the writing test in the school exam also proves that there is an attempt to improve the quality of the English school exam. However, the implementation of the school exam is not always in line with the expectation since there are still some weaknesses. Firstly, most teachers are in doubt about their expertise in "constructing the high-quality items and accuracy in interpreting students' score derived from the test when they are granted autonomy in evaluating their students learning" (Sulistyo, 2015).

Also, based on the analysis of the sample of items for the school exam conducted in one school in Semarang, Central Java, there are some weaknesses of the test items that could affect the validity and reliability of the test. As an example, question no. 45 (see image p. 3) could possibly cause ambiguity for students. The aim of item no. 45 based on the test indicators is to elicit a conclusion. However, the item does not provide enough clues for students to conclude. The answer to such a question could be varied so it is difficult to give a score objectively. The fact that an individual candidate might interpret the questions in different ways and different occasions means that the item does not meet the principles of test reliability (Terry & Hughes, 2006).



Read the text and answer questions 45.

Dear Anne

You are invited to
A graduation Party
Honoring
Sophia Jenkins

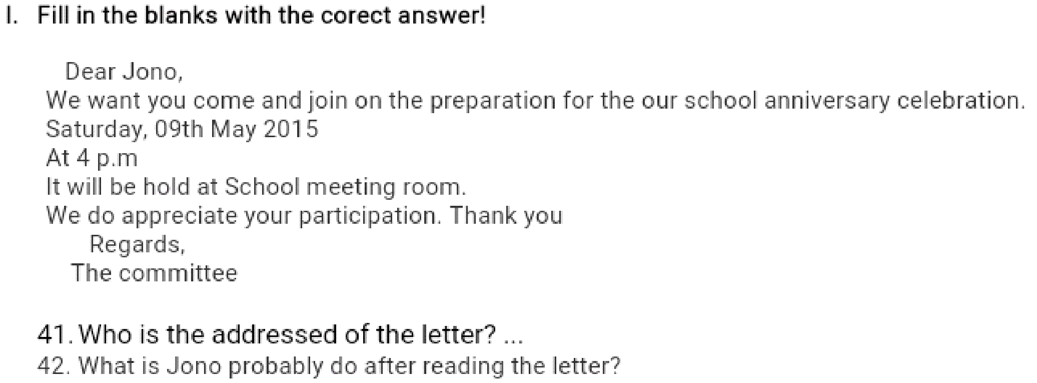
Saturday, May 17
6:00 – 9:00 PM
The Jenkins Residence
402 Northside – ONVE
NASHVILLE, TN 37925

Best Wishes
- The Jenkins -

45. What can we conclude from the dialog?

Taken from (Saukah, Cahyono, & Cahyono, 2015) National Standard School Exam, English at Junior High School in the academic year 2017/2018

Another example is item no. 41 is mistaken in using the appropriate noun. Rather than the addressed, the addressee is the appropriate one. Additionally, item no. 42 is grammatically incorrect since the auxiliary verb for *do* is not “is” but the auxiliary verb for the verb do is “does”. Also, item no. 43 also misses the auxiliary verb.



I. Fill in the blanks with the corect answer!

Dear Jono,
We want you come and join on the preparation for the our school anniversary celebration.
Saturday, 09th May 2015
At 4 p.m
It will be hold at School meeting room.
We do appreciate your participation. Thank you
Regards,
The committee

41. Who is the addressed of the letter? ...
42. What is Jono probably do after reading the letter?

Taken from (Saukah et al., 2015) [National Standard School Exam, English at Junior High School in the academic year 2017/2018]

Based on some mistakes in writing items for the school exam, it can be understood that writing good quality items is still a concern for teachers. This also supports the previous research finding asserting that most locally produced achievement tests are not designed satisfactorily (*Azwandi (1996) referensi dari tesis.pdf*, n.d.). Even though the examples above cannot be generalized, the level of validity and reliability of the test can decrease due to some mistakes in writing test items. Notwithstanding the weakness of items for school exams, the content validity of the school exam is good since the test items are written based on specifications designed by *MGMP*. They understand the scope of materials in the English curriculum precisely. Nevertheless, the shortcomings of the school exam for Junior High School is the absence of the writing test. Despite the inclusion of open-ended questions, the questions only measure students' reading understanding.

Considering that there are still some shortcomings of tests made by teachers, this article provides an example of test design that could be used as considerations for English teachers in designing English test. The test design was also administered in a Junior High School involving ten students to take the test voluntarily. The results of the test then were analyzed and evaluated. The recommendations for test improvement were also made by considering the results of the test evaluation.

Research Methodology

The testing context

The test is called a midterm test (MT) occurring at the end of the 12th week of a 26-week course. This test could be categorized as a progress test since it aims to measure the students' language and skill progress in relation to the syllabus they have been following (Harmer, 2004). This MT focuses on reading skills. According to the school-based curriculum in Indonesia, these skills become the focus of the English curriculum for junior high schools (Lie, 2007). Also, familiarizing the students with the test items of the national exam is important because the reading test is commonly presented to junior high school students in Indonesia. The target audience of this test is junior high school students in year 9 (aged between 14 and 16 years) at Semarang-Central Java. According to my colleague who teaches English at this level in Indonesia, there are different levels of proficiency among the students. Most of them are at a beginner level but some are already at a pre-intermediate level (M. Muttaqien, personal communication, October 10, 2018). This MT is designed following the standard and basic competence of English predetermined by the 2006 school-based curriculum in Indonesia because the syllabus is a guideline for teachers in achieving the objectives of the English instructions. The syllabus is designed to teach the students the meaning of short genre texts in the form of procedures, reports, narratives and some short functional texts (such as brochure and personal letter) (Indonesian Ministry of Education and Culture, 2006). More specifically, the students are encouraged to develop reading skills, such as identifying the main ideas of a paragraph, specific information, pronominal reference, contextual meaning of a word, and communicative purposes of the texts (Indonesian Ministry of Education and Culture, 2006).

Purposes of the test

This MT aims to recognize students' preliminary achievement of the basic competence of reading, predetermined by the English curriculum, at the half-way point of the 6-month semester. The MT is also aimed at preparing the students to practice the national examination that will be conducted at the end of the semester. Additionally, the result of this MT will be taken into consideration when determining or allocating the English teacher's performance results at the half-way point of the semester. Furthermore, the results of the MT will be provided as individualized feedback for the students and help the teachers identify those students who need additional assistance. Lastly, the purpose of the MT is ultimately to make the students aware of the importance of reading skills in a real-life because reading skills are important in order for the students to communicate within an environment where English is used as a means of communication. Even though the students do not live in an English-speaking country, such as Indonesia, it is easy for them to find reading materials written in English, including newspapers, books, and internet resources.

Possible difficulties

There were several problems I encountered when designing this MT. Firstly, deciding on the appropriate texts is difficult because of the different levels of the students' English proficiency, that is a beginner and pre-intermediate level. Considering the level of text difficulty and the students' proficiency level is imperative because if the texts are too difficult, the MT cannot measure the students' reading ability in their level and can possibly make them frustrated. Secondly, the Indonesian government has high expectations to generate graduates who have higher order thinking skills that do not comply with students' English competence. To overcome this problem, I needed to design an MT that could follow the criteria of the syllabus and try to incorporate a few higher order thinking skills. I also had trouble to find the authentic texts that could meet the reading operations that will be exercised in the MT. This is because not all authentic texts could be used as test materials. To overcome this problem, I tried to modify the texts to meet the intended purpose for the MT.

Text types

Regarding the test materials, the reading texts used for this test are derived from authentic texts, such as a manual, a brochure, a report text, and short folk tales. The authentic text from real-world sources is chosen as an attempt to meet the authenticity of the test (Brown & Abeywickrama, 2010). The length of each text ranges between 150-400 words. When choosing the texts, I also considered the readability index of each text that can be found in Microsoft Word under the *File > Option> Proofing>show readability statistics* function. The details of the readability index of the chosen texts are as follows. The first text is a procedure text entitled *How to make a balloon-powered bath boat*. The readability index of this text based on the readability statistic accounts for 85.6 for Flesch Reading Ease (FRE) and 6.0 for Flesch-Kincaid Grade Level (FKGL). This readability index shows that the higher the score of FRE, the easier the text will be. In contrast, the higher the FKGL, the more difficult the text will be read. The second text is an Indian folktale with reading index FRE 78.4 and FKGL 5.2. The third text is a report text about Carpenter ants. This short report text has FRE 65.4 and FKGL 7.8. The fourth text is a brochure about a tourist resort in Bali. The readability of this text is low since it has FRE 53.7 and FKGL 10.6. Even though, this text is considered a difficult text, it is still useable. According to (Nielsen, 2011), if the readability index of a text is under 11th grade-reading level, the reading texts can be read by high school students, under year 11. The last text is a fable about a monkey and a crocodile which has a readability index FRE 85.2 and FKGL 4.1. In sum, the texts that will be used in the MT have met the criteria of readability, and they can be used to test the candidates' reading comprehension.

The reading texts are also selected by considering the background knowledge of the candidates. I tried to choose the texts which are not too familiar to the candidates' background knowledge. This aims to prevent the candidates from answering the questions without reading the text, which may occur if the students already were familiar with the content of the texts (*(Cambridge handbooks for language teachers) Arthur Hughes-Testing for Language Teachers-Cambridge University Press (1989).pdf*, n.d.). I also avoided choosing texts which were culturally laden as suggested by Hughes (2003) because this test is only to measure candidates' reading ability. Finally, I choose texts that had not been read by the candidates because familiarity with the texts could aid the candidates when answering the questions and could decrease the validity of the test.

Test tasks

The test tasks that I developed constitute interactive reading tasks since there is a combination of form-focused and meaning-focused objective, but the meaning is emphasized more Brown & Abeywickrama, 2010). This is because the purpose of the reading test that I developed, that is, the candidates are encouraged to be able to understand the meaning of the texts. Additionally, there are four reading tasks in this MT. They are impromptu reading plus comprehension questions, cloze tasks, short-answer tasks, and open-ended reading comprehension questions (Brown & Abeywickrama, 2010). The reading types employed in this MT include expeditious and careful reading (Urquhart, 1998). I also incorporate global and local reading (Urquhart & Weir, 1998) to develop both weaker and stronger students' reading skills.

As an attempt to translate standard competence of reading skills determined in the syllabus, this MT adopts some reading operations from literature, including (i) skimming quickly to establish discourse topic and main ideas; (ii) searched reading to locate quickly and understand information relevant to predetermined needs; (iii) reading carefully to establish accurate comprehension of the explicitly stated main ideas the author wishes to convey propositional inferencing; (iv) guess meaning of an unknown word based on context; and (v) identify pronominal reference (Brown & Abeywickrama, 2010; Hughes, 2003). These skills were chosen because I considered that both beginner and pre-intermediate level could achieve the aforementioned skills. Furthermore, recalling to the purpose of this MT, I want to recognize which reading skill that students need to improve. The specifications of this MT are explained in the following paragraph.

In Part 1, multiple-choice questions (MCQs) are used. I realized that MCQs have some weaknesses and get criticized (Terry & Hughes, 2006); however, in the sense of practicality, I argue that this test task is still worthwhile to be included in the MT. This is because classes in Indonesian schools are big. Overall, the number of students in a classroom could reach 40-50 students and teachers usually teach two or three classrooms. Additionally, Hughes (2003) also contends that even though MCQs have some weaknesses, it does not mean MCQs must be eliminated from the test. Further, he argues that "in reading a comprehension test there may be certain tasks that lend themselves very readily to the multiple-choice format" (p. 78). Accordingly, I still want to employ MCQs for the MT by considering their practicality and keeping in mind suggestions from Hughes (2003), such as avoiding being excessive, indiscriminate and potentially harmful use of the technique.

In Part 1, the candidates will encounter a procedure text and a story. They will be instructed to answer ten questions in 15 minutes. One example of an item in the test is that students will make an inference about the benefits of reading a procedure text about how to make a balloon-powered bath boat. In answering this item, the candidates are encouraged to employ their skills in making an inference beyond the sentence level.

In Part 2, a cloze test is used to measure how well the candidates' reading comprehensions are (Nielsen, 2011). In the cloze task, the candidates need to fully understand the main idea and specific information provided in the text; thus, careful reading will be exercised here. Unlike multiple-choice tasks, in this cloze task, the candidates can not solely rely on the guessing technique since the distractors of the cloze task are more than in the multiple-choice tasks. In this part, the candidates will read the report text and fill in the blank of the summary of the report text to answer questions from 11 to 15. The procedure of this section is that the candidates will fill in an incomplete summary based on the text with appropriate word lists provided in the test sheet. To accomplish the cloze task, the candidates need to identify important ideas and specific information provided in the text. Also, they need to employ reading strategies, such as skimming and scanning. In this section, the candidates will spend 10 minutes to complete the questions.

Part 3 employs a short answer task. This test task aims to measure candidates' skills in identifying the addressee or audience of the text and scanning for a specific name. The short answer task is administered to measure candidates' comprehension by writing the answer instead of choosing the available answers like in the multiple-choice task. Urquhart and Weir (1998) argued that reading comprehension should be achieved by answering a correct answer rather than based on candidates' test-taking strategies (such as guessing or matching). Since the candidates are encouraged to employ expeditious reading by scanning and skimming the passage, this part allocates 5 minutes only.

In Part 4 there are three open-ended reading comprehension questions. The objective of this task is to enable the candidates to make propositional explanatory inferences or answering questions beginning with *why* and *how* (Hedges, 2000), and formulating the moral values of the story. This test task could possibly lower the reliability of the test since there will be various responses from the candidates. This is confirmed by Brown and Abeywickrama (2010) who assert that open-ended reading comprehension questions could possibly pose a problem in reaching intended criterion. This is because creating consistent specifications for acceptable students' responses is difficult (Brown & Abeywickrama, 2010). However, the disadvantages may be outweighed by some benefits of this test task, such as enabling candidates to construct their own answers and promoting follow-up discussion (Brown & Abeywickrama, 2010).

The scoring scheme and/or relevant criteria

To be able to measure the candidates' achievement in the reading test, the scoring method needs to close attention to detail. This is because the scoring method is an important part of the measurement process (Bachman, 1996). Further, the emphasis on the scoring of language testing is crucial because it can be taken into consideration in deciding (Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, 1984). For example, in this test, the score will be used by the teacher in assisting the weaker students and adjusting teaching practice. Therefore, the scoring criteria and procedures of marking the students' responses to the test tasks need to be formulated thoroughly by taking into consideration theories from the literature, such as validity and reliability.

The criteria for scoring in this MT are as follow. In Part 1, for multiple-choice questions, in each item, there will be only one correct answer out of four options. This is because multiple-choice questions are categorized as a selected response task and should be scored objectively. In part 2, for the cloze task, there are five gaps that must be filled in with 5 correct answers from seven options. Misspellings in transferring the answer will be considered wrong since there are provided word lists in completing this test tasks. In Part 3 and 4 where there is a short-answer task and an open-ended reading comprehension task, the scoring criteria are a bit different since the candidates are required to write their answers. When the candidates make grammatical errors or misspellings in responding to the questions, their score will not be deducted. This is because the main purpose of the reading test is to measure candidates' reading ability. The grammatical errors and misspellings are not considered as a serious issue in determining the success of candidates' performance on the tasks (Hughes, 2003). In addition, considering that part 4 is open-ended reading comprehension questions, the variety of students' responses will be considered. For example, in the questions about moral values and text addressee, students' responses may vary. There may be an issue related to the rater's subjectivity in giving the score and could lower the validity of the test. To overcome the problem, I will follow (Brown & Abeywickrama, 2010) advice by creating clear instructions (such as specifying the word number and/or sentence in the test sheet) and making a uniform rubric for scoring. To assist with consistent scoring, there will be three criteria of the candidates' answers, including completely wrong, partially right and completely right answers. Additionally, to obtain objectivity, there will be two scorers in this test: the test developer and the classroom teacher. The aim of employing two scorers is to obtain inter-rater reliability (Brown & Abeywickrama, 2010).

The score for a correct answer in multiple-choice questions and the cloze task is 1 and for a wrong answer is 0 as suggested by Bachman and Palmer (Bachman, 1996) In short-answer questions, the scoring technique is the same with multiple-choice questions: the correct answer will be valued 1 point and the wrong answer will be valued 0. In open-ended reading comprehension questions, a completely wrong answer will be awarded 0, a partially correct answer will be awarded 1 and a completely correct answer will be valued 2 points. The details of the points in each part are as follows. In Part 1, there are ten items; therefore, the point will be 10. In Part 2, there are five items, the point will be 5. In Part 3, there are only two items with one point in each item; therefore, the point will be 2. In Part 4, there are three items with two points for a completely correct answer in each item. In Part 4, more points are given because the students are encouraged to employ higher order thinking skills, such as making proportional explanatory inferences and they are also instructed to write their own answers. The candidates who can answer three questions correctly will get 6 points. In total, the candidates who answer all questions correctly will get 23 points. The candidates who get 60 in the MT, will have achieved the minimum passing criteria determined by the school (M. Muttaqien, personal communication, October 14, 2018). For those who get below 60, they need extra support in the second half of the semester.

Administering the test

After finishing the test design and writing the test scripts as well as reviewing the test scripts with a native English speaker, this test was administered in an Islamic junior high school in Indonesia. The test scripts were sent to my colleague in Indonesia who teaches English at the Islamic junior high school in Semarang, Central Java. There were ten volunteers participating in this test.

What follows is the explanations of my test, encompassing (1) the degree of communicativeness and innovativeness of the test, the rationale of choosing the test design, (2) potential backwash of the test, (3) practical rationale for the design, (4) the presentation of test results with some comments, and (5) the evaluation for improvements of the test.

Findings and Discussion

The significant features of the test: communicativeness and innovativeness

According to Morrow (Morrow, 2018), "communicative language testing was a movement to enhance the validity of language tests by incorporating authentic materials and activities based on predictions of the testees' target language use" (p. 1). As an attempt to achieve communicativeness of this MT, I used some authentic texts that students often encounter in daily life. For example, I selected texts from various genres, including procedure text, fables, brochures, and folk tales. These texts are often found in students' daily life. For example, when they read magazines they find recipes, manuals, instructions and so forth. Additionally, in Part 1, I provide a procedure text on how to make a balloon-powered bath boat. This text is appropriate for the level of their ages (adolescent) to make them interested in reading the texts. Hughes (2003) confirms that in selecting reading texts, test developers need to "choose texts that will interest candidates, but which will not overexcite or disturb them" (p. 142).

Other than that, the communicativeness of this MT could be found from the reading tasks that enable candidates to interact with the texts and the author. This is because, according to (Hedges, 2000) reading is an interactive process, where there is "a dialogue between the reader and the text or even between the reader and the author" (p. 188). As an example, in the test papers, there is a question asking about the text addressee and the benefits of reading the text. This means the test asks students to think about the relation between the text they read and real life.

Regarding the innovativeness, I admit that I used old fashioned test items, such as multiple-choice questions, a cloze task, and short answer questions that follow the passage. Even though the test items are considered old fashioned, I tried to make it become innovative in some ways. Firstly, I made the questions related to students' real life. For example, I include questions that require students to think about the benefits of reading the texts. Secondly, I also included items that require students to connect between the content of the text and their experience. More specifically, questions about the moral values of a certain story could be varied so that it could provoke students and teachers discuss the various possible answers. Furthermore, the innovativeness of this test also could be seen from the layout of the test. I put pictures to set the scene of the test. Not all parts have pictures, but at least I have tried to make the test more interesting. Students not only read the texts, but they also have images to help them understand the context and stimulate their schemata.

Additionally, open-ended reading comprehension questions could be categorized as an innovative part of this test. This test task could possibly lower the reliability of the test due to various responses that the candidates can make. However, the disadvantages may be outweighed by some benefits of this test task, such as enabling candidates to construct their own answers and promoting follow-up discussion (Urquhart, 1998). A follow-up discussion is important since it could stimulate students to think analytically and critically. It could also motivate students to learn autonomously since they are curious to find the correct answer.

Potential for beneficial washback

According to (Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, 1984), in the classroom testing, washback constitutes "information that 'washes back' to the students in the form of useful diagnoses strengths and weaknesses" (p. 29). The MT also has the potential to recognize students' strengths and weaknesses in terms of their reading skills. As an example, from analyzing the test results, the strong and weak students can be recognized. More specifically, it is found that the weak students are still unable to construct their answers in open-ended reading tasks (Part 4). Most of them cannot summarize the points in the texts and write the answers with their own language. Most of them know the location of the answer but they were not successful in transferring the information with their own understanding. Knowing this, then English teachers in junior high school need to introduce and train the skills of summarizing and paraphrasing.

More specifically, in Part 4, I included open-ended questions to measure students' ability to find the main idea in a paragraph of the text. To this end, I chose one paragraph where the main idea is not stated in the first sentence of the paragraph. Instead, I chose a paragraph that has the main idea, but students should summarize it from the paragraph. This is a rather difficult question since it needs summarizing skills that cannot be obtained from doing multiple-choice items. From the test, it is found that students still face difficulty to answer such items. The backwash of this item test is that teachers need to teach students how to make a summary. This is very important for academic purposes since they are required to write an essay or a thesis later in a higher level of education. They need to learn skills for making a summary from the junior high school level in order to be skillful in the higher level of education, at the university for example.

A practical rationale for the design

In choosing the test design I try to make it balanced between multiple-choices questions and other test items. Often students have many multiple-choice items. For example, in the national exam, out of 50 questions, 45 are multiple choices. However, in this test, I tried to make the proportion of the items be balanced between multiple choices and other test items. In fact, there are ten multiple choices, five clozes, and ten short answer questions.

Additionally, as an attempt to translate standard competence of reading skills determined in the syllabus, this MT adopts some reading operations from the literature, including (i) skimming quickly to establish discourse topic and main ideas; (ii) searched reading to locate quickly and understand information relevant to predetermined needs; (iii) reading carefully to establish accurate comprehension of the explicitly stated main ideas the author wishes to convey propositional inferencing; (iv) guess meaning of an unknown word based on context; and (v) identify pronominal reference (Brown & Abeywickrama, 2010; Hughes, 2003). These skills were chosen because I considered both beginner and pre-intermediate levels should achieve the skills. Furthermore, recalling the purpose of this MT, I want to recognize which reading skills students need to improve.

Another aspect that I want to explain is the layout of the test sheets. I put a translation of the instructions of the test sheet into Bahasa Indonesia to make students more understand how to answer the questions. This is because I don’t want to make the reliability of the test low if students did not understand the instructions of the test. For example, at the beginning of the test sheet, I put general instruction. I also put specific information for each part, such as the maximum words should be written in part 3 and several sentences, they need to respond to questions in Part 4.

In addition, in each part there is also time allocation to prepare in completing the test and students can maximize the time provided. Also, in each part, there is a weighted point to make students put their effort into each part. Harris and McCann

(1994) maintain that "it is very useful to write the marking scheme on the test paper itself so that the students also know how much each section is worth" (p. 56).

Presentation of the test result

There are four parts in this test, including multiple-choice questions (Part 1), close tasks (Part 2), short answer questions (Part 3) and open-ended reading questions (Part 4). Unlike Parts 1, 2, and 3, in Part 4, I give a different point for each correct answer, that is, two points for each correct answer. The reason is that in Part 4 students are advised to write their own answers in one sentence. Part 4 also requires students to employ skills of making propositional inferences, which are difficult. In part 4, I also did not employ statistical calculation as I did in parts 1, 2, and 3 because Part 4 has a different score for each correct answer and it is open-ended questions. However, the results of Part 4 could be used to validate students' score in the previous parts. Also, Part 4 can be used as the basis of making decisions about weak and stronger candidates. This reflected in the students' answers since in Part 4 students are required to construct their own answer by writing it with their own words. The results of the test can be seen in the table below.

Table 1. Test results

Students	Total score from parts 1 to 3	Part 4
S1	14	2
S2	15	2
S3	13	2
S4	10	0
S5	5	0
S6	17	5
S7	19	10
S8	10	0
S9	19	7
S10	11	2
Mean	13.3	-
SD	4.22	-

Based on the statistical analysis (Appendix 1), it is found that there are some weaknesses in my test items since there are several items that have very high facility values (FV) and very low discrimination index (DI) as well as negative DI. Firstly, the facility values of the items in Part 1, 2, and 3 range between 0.3 and 1.0. Based on FV, the number of easy and difficult questions in this test is a balance. There are

10 questions that are categorized as easy since the FV ranges between 0.7 and 1.0 and there are ten questions that are categorized as difficult since the FV ranges between 0.3 and 0.6. According to (Saukah et al., 2015), the items which have more than 0.66 is categorized as easy items and need to be revised or replaced. However, not all items that have high FV should be dropped. This is because the easy items could motivate weak students to attempt to complete the test (McNamara, 2000). Furthermore, I have mentioned in my test design that the students taking the test have diverse proficiency. I do not want to make students feel frustrated because all the questions are so difficult. This test is also not a proficiency test that requires students to achieve a predetermined passing grade. Instead, this test is a progressive test in the midterm. Hughes (2003) also confirms that not all items with high facility values must be eliminated from the test.

Regarding the DI, there are some questions with very low DI even negative. In fact, the DI of this test ranges between -0.2 and 0.8. According to Heaton (*Heaton pp.88-134.pdf*, n.d.), an item with a negative discrimination index is highly inadvisable to be used again since it gives the wrong direction. The strong candidates do badly on that item and vice versa.

Relating to the standard deviation, this test has a low standard deviation (4.22). It means that the spread of the students' score is small. Heaton (Heaton, 1988) confirms that "if the test is simply to determine which students have mastered a particular program of work or are capable of carrying tasks in the target language, a standard deviation of 4.08 or any other denoting a fairly narrow spread will be quite satisfactory provided it is associated with a high average score" (p. 176-178). This is in line with the purpose the MT which is simply to recognize students who have mastered reading skills based on the predetermined specifications of the syllabus and three months period of learning English in a six months semester course.

Evaluation of the test design

According to Brown and Abeywickrama (2010), there are at least five principles of language assessment, including practicality, validity, reliability, washback, and authenticity. In this section, the four principles are used as the basis to discuss and evaluate the MT. This is because the principle of washback has been discussed above.

The validity of this test will be discussed first. According to Harmer (Jeremy Harmer, 2003), a test is valid if it tests what it is supposed to test. This kind of validity is called construct validity. This MT is categorized as reading test that measures students' reading skills, including identifying main ideas of texts, exercising students' reading skills, including skimming, scanning and inferring the meaning of words based on contexts. Students are also advised to read the texts and answer some questions related to the texts to measure students' comprehension.

Following the definition of validity explained by Harmer (2015), it can be inferred that this test has met the construct validity.

Validity also has another variation called content validity. According to Hughes (2003), a test is considered to have content validity "if its content constitutes a representative sample of language skills, structure, etc." (p. 26). This MT is also designed by incorporating some operations of reading suggested by Hughes (2003) and following the content of the syllabus prescribed by the English curriculum in Indonesia. The text types used in this test is also in accordance with the text types suggested by the syllabus. In terms of content validity, it can be inferred that this MT has met the content validity.

What is more, this test has met the face validity based on the number and time allocation to complete the test. In fact, there are 25 questions that must be answered by students in 60 minutes. In average, they will answer each question within 2.4 minutes. Based on my discussion with my colleague who teaches the class, the time allocation in this test is reasonable and fits with students' English ability (Z. Muttaqien, personal communication, October 18, 2018). The texts used in this MT also have been graded based on the level of the readability by using a default function in Microsoft words to grade the readability of the texts. The statistical analysis of the function in Microsoft words also has suggested that the readability of the texts is appropriate.

The second principle of language assessment is reliability. According to (Harmer, 2004) "Reliability refers to the consistency of the test results. Given the same conditions, a test should be always given the same results" (p. 409). In achieving test reliability Hughes (2003) suggest test developers provide clear and explicit instructions. In this test, the clear instructions have been provided. There are general instruction and specific information for each part of the test. There is also the Indonesian translation to make students understand the instruction easier.

Additionally, Hughes (2003), to gain reliability, a test should provide a detailed scoring key. This MT also has a scoring guide that I have explained earlier. Recalling the information provided in the previous assignment, the test is scoring objectively by considering the right and wrong answer students give. The correct answer is awarded one point and wrong answer zero. The scoring guidance is for parts 1, 2 and 3. The grammatical errors are not included as part of the scoring guide as long as students can answer the correct answer they will be given one point for each correct answer. Unlike Parts 1, 2, and 3, in Part 4 there is a different scoring guide since, in this part, students are advised to write their answer in a sentence and paraphrasing their answers. Therefore, this part has more points, that is, two points for each correct answer. Heaton (1988) confirms that when marking open-ended items which require answers in a sentence, it is frequently advisable to award at least two or three marks for each correct answer. In the MT, two marks are given for each correct answer. The

Heaton's (1988, p. 133) recommendation in marking open-ended items is also adapted as follows.

- Score 2 is for a correct answer in a grammatically correct or a sentence contain a minor error.
- Score 1.5 is for a correct answer in a sentence containing one or two minor errors (but causing no difficulty in understanding)
- Score 1 is for a correct answer but very difficult to understand because of one or more major grammatical errors
- Score 0 is for an incorrect answer in a sentence with or without errors.

The last principle of language assessment is practicality. According to Brown and Abeywickrama (2010), an effective test is practical if it "is not excessively expensive; stays within appropriate time constraints; is relatively easy to administer; has scoring/evaluation procedure that is specific and time-efficient" (p. 19). Some of the aspects of practicality have been discussed earlier, one aspect that needs to be explained is the administration of this test. The MT is very easy to be administered by paper and pencil. It is also easy to mark since the key answers and some marking criteria to decide a correct and wrong answer have been prepared.

Conclusion and Pedagogical Implications

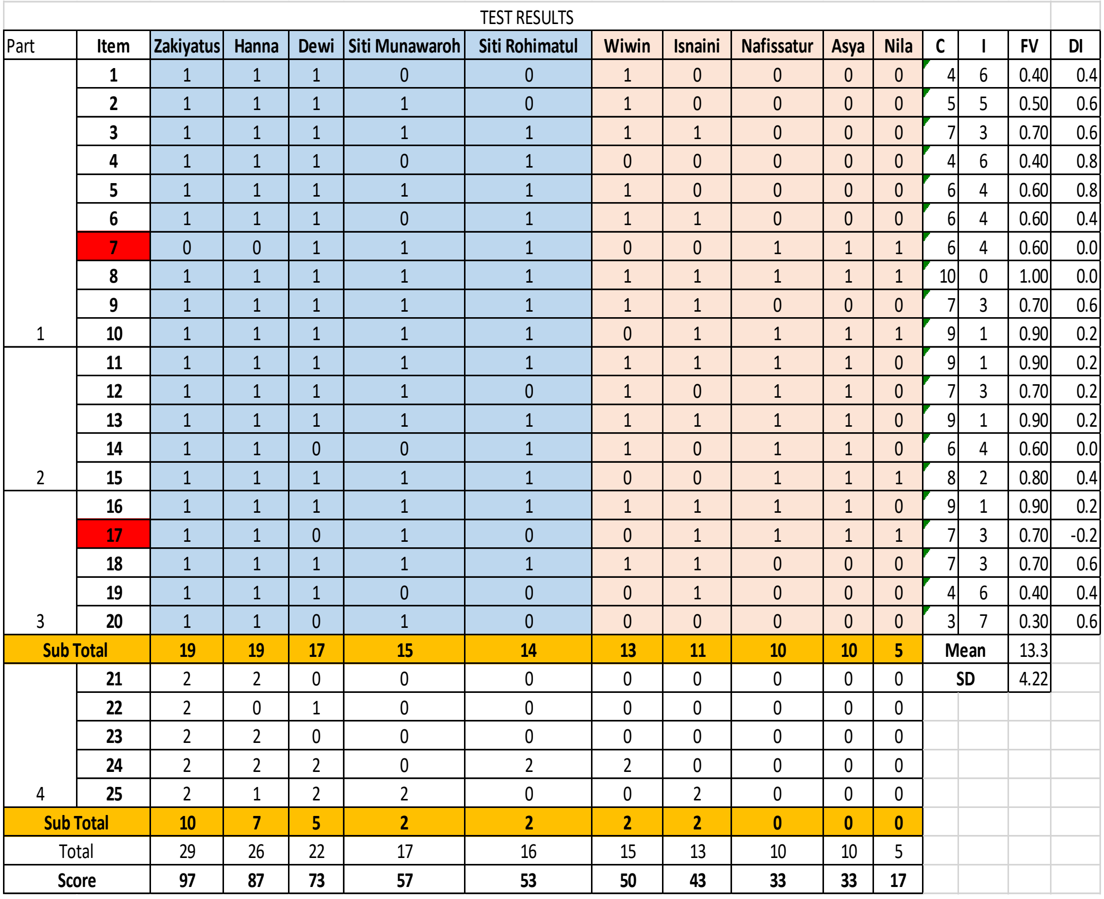
Based on statistical analysis, there are several items that have very low facility values. As an attempt to improve the test, some questions with very low facility values should be revised and/or dropped. It is also necessary to change the text for Part 3 since the text is too difficult and the structure of the sentences are too complex so that there are some strong students could not answer the items in Part 3.

Item number 7 Part 1 is needed to be revised. This is because the item does not discriminate the strong and weak candidates. When it was checked, there was ambiguity in the multiple choices. In fact, even though, the facility value of the item is not too difficult (0.60), the strong candidates still could not answer the question. Therefore, there is something wrong with the distractors of the item (See Appendix 2).

It was also found that many students could not answer the questions in Part 4 since the test requires students to summarize the main idea from one of the paragraphs and paraphrase their answers. However, most of them wrote their answers by picking up one or two sentences from the passage in the first or the last sentence. It seems that the students follow a rule of thumb that the main idea is always located in the first or last sentence in each paragraph so that they are trapped by the rule of thumb. They also could not summarize their answers in one sentence; instead, they wrote their answer in more than one sentence. Knowing this fact, the English teachers need to teach students how to make a summary and to paraphrase sentences.

To conclude designing a test is not easy. Test designers need to consider aspects of testing and assessment. In designing this test, the author also finds some challenging issues as it is mentioned earlier. Although many efforts have been put forward, this test design also has some weaknesses that need to be refined. Therefore, for the next test designers, some weaknesses in this test can be used as consideration. Additionally, this test design is for a small scale, the future researchers are encouraged to design a test for large scales, such as a national exam.

Appendix 1



TEST RESULTS

Part	Item	Zakiyatus	Hanna	Dewi	Siti Munawaroh	Siti Rohimatul	Wiwin	Isnaini	Nafissatur	Asya	Nila	C	I	FV	DI
1	1	1	1	1	0	0	1	0	0	0	0	4	6	0.40	0.4
	2	1	1	1	1	0	1	0	0	0	0	5	5	0.50	0.6
	3	1	1	1	1	1	1	1	0	0	0	7	3	0.70	0.6
	4	1	1	1	0	1	0	0	0	0	0	4	6	0.40	0.8
	5	1	1	1	1	1	1	0	0	0	0	6	4	0.60	0.8
	6	1	1	1	0	1	1	1	0	0	0	6	4	0.60	0.4
	7	0	0	1	1	1	0	0	1	1	1	6	4	0.60	0.0
	8	1	1	1	1	1	1	1	1	1	1	10	0	1.00	0.0
	9	1	1	1	1	1	1	1	0	0	0	7	3	0.70	0.6
	10	1	1	1	1	1	0	1	1	1	1	9	1	0.90	0.2
2	11	1	1	1	1	1	1	1	1	1	0	9	1	0.90	0.2
	12	1	1	1	1	0	1	0	1	1	0	7	3	0.70	0.2
	13	1	1	1	1	1	1	1	1	1	0	9	1	0.90	0.2
	14	1	1	0	0	1	1	0	1	1	0	6	4	0.60	0.0
	15	1	1	1	1	1	0	0	1	1	1	8	2	0.80	0.4
3	16	1	1	1	1	1	1	1	1	1	0	9	1	0.90	0.2
	17	1	1	0	1	0	0	1	1	1	1	7	3	0.70	-0.2
	18	1	1	1	1	1	1	1	0	0	0	7	3	0.70	0.6
	19	1	1	1	0	0	0	1	0	0	0	4	6	0.40	0.4
	20	1	1	0	1	0	0	0	0	0	0	3	7	0.30	0.6
Sub Total		**19**	**19**	**17**	**15**	**14**	**13**	**11**	**10**	**10**	**5**	**Mean**		13.3	
4	21	2	2	0	0	0	0	0	0	0	0	**SD**		4.22	
	22	2	0	1	0	0	0	0	0	0	0				
	23	2	2	0	0	0	0	0	0	0	0				
	24	2	2	2	0	2	2	0	0	0	0				
	25	2	1	2	2	0	0	2	0	0	0				
Sub Total		**10**	**7**	**5**	**2**	**2**	**2**	**2**	**0**	**0**	**0**				
Total		29	26	22	17	16	15	13	10	10	5				
Score		**97**	**87**	**73**	**57**	**53**	**50**	**43**	**33**	**33**	**17**				

Appendix 2

READING TEST

Mata Pelajaran	: Bahasa Inggris
Kelas	: IX (Sembilan)
Hari/Tanggal	: Kamis, 25 Oktober 2018
Durasi	: 60 Menit

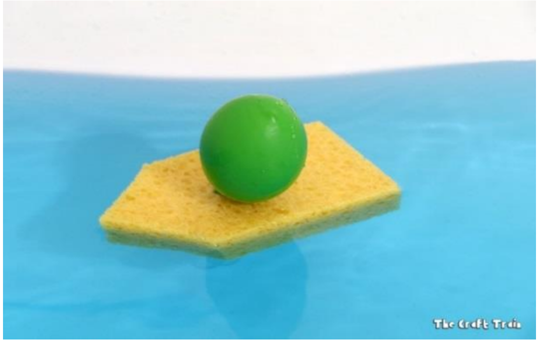
PETUNJUK UMUM

1. Bacalah dengan teliti petunjuk dan cara pengerjaan;
2. Kerjakan pada lembar jawab yang tersedia;
3. Tidak diperbolehkan menggunakan kamus;
4. Periksalah kembali seluruh pekerjaan sebelum diserahkan kepada bapak/ibu guru Anda.

PART 1 (10 points)

Read the text and answer the following questions by choosing the best answer (a, b, c, or d). You should spend 15 minutes in this section. Text 1 is for question number 1-5

How to make a balloon-powered bath boat



You will need:

- 1 x thick sponge
- 1 x balloon
- 1 x piece of 2cm clear plastic tubing
- craft knife
- scissors

Stages

1. Use the scissors to cut a pointed end to the sponge to make the bow of your boat.
2. Now in the stern of the boat, use the craft knife to make a small slit. Feed the plastic tubing through the slit so that it sits halfway.
3. Blow up your balloon and let the air out to make the balloon nice and stretchy.
4. Now slip the end of your balloon over the top of the plastic tubing - the bit sticking through to the top of the boat.

5. Fill up the bathtub and blow the balloon up by placing your mouth over the lower part of the plastic tube - the bit sticking below the boat's 'hull'. Place a finger over the end of the tube to stop the air coming out straight away.
6. Put your boat into the water and remove your finger, the rush of air from the deflating balloon should propel your boat across the bath.
7. You'll have to squeeze the water out of the sponge each time you launch your boat.

Questions

1. Which of the following is **NOT** one of the benefits of reading the passage?
   a. Knowing how to launch a balloon-powered bath boat
   b. Making a balloon-powered bath boat
   c. Getting inspiration to make a craft
   d. Knowing how to blow up a balloon

2. What must readers do with the tube to stop the air coming out?
   a. Remove a finger from the end of the tube
   b. Block the end of the tube with finger
   c. Keep blowing the tube
   d. Put the tube into the water

3. What must readers do every time the boat is launched?
   a. Squeezing the water
   b. Removing your finger
   c. Stopping the air
   d. Propelling the boat

4. What does the word “it” (stage 2) refer to…
   a. The boat
   b. The craft knife
   c. The plastic tubing
   d. The slit

5. The word “stretchy” has a similar meaning with…
   a. Hard
   b. Flexible
   c. Stiff
   d. Sticky

Text 2 is for question number 6-10

Many years ago, it is said, that elephants had small trunks with stubbed noses. One year, it did not rain for many months. The ponds and lakes began to dry up, and the streams had very little water. All the animals in the forest were very thirsty, and desperately searching for a source of water. A river used to flow not very far away from the forest, and an elephant decided to go there in search of water.

Walking slowly, he reached the river. There lived a bright green crocodile in the river. As he saw the elephant, he cried, "Go away! Water is already scarce here. If you start drinking, what will be left for me?" The elephant knew it was a risk to pick a fight with the crocodile. So, he decided to come back to the river when the crocodile would be sleeping.

In the same river, there also lived a shiny green toad. Whenever the crocodile would be swimming across the river, the toad would hop onto his back and enjoy a ride. Over time, the crocodile was annoyed with giving free rides to the toad. Many times, he had tried to shake the toad off his back, but in vain. "Hahaha!" the toad would laugh.

One day, the crocodile was resting on a rock. Finding this to be a good opportunity, the elephant went to the river silently and began to drink water. Just then, the toad jumped onto the crocodile's back, disturbing his slumber. The crocodile was irritated! He began to swim around the river and shake his body violently. "Now, I shall get rid of you!" he cried at the toad. But, the toad was unmoved.

Suddenly, the crocodile noticed the elephant. "How dare you drink from my river when you were told not to?" he cried. Unable to get rid of the toad, the crocodile decided to vent all his anger on the elephant. He caught the elephant's trunk and began to pull him into the river. The poor elephant started to pull back, crying, "Let go of me....please! Let go of me....my nose hurts!" But the crocodile showed no mercy. Then, with a mighty jerk, the elephant succeeded in freeing his trunk from the crocodile. But, in tug of war, the elephant's nose had become really long! Angry, the elephant sucked all the water from the river. Then, he sucked some mud and sprayed it on the crocodile and the toad. Since then, it is said, elephants have had long trunks, and crocodiles and toads are not bright green anymore.

Questions

6. The text tells us about….
   a. Crocodile, elephant and toad
   b. The origin of elephant’s trunk, and the colour of crocodiles and toads
   c. The elephant and crocodile are looking for sources of water
   d. A conflict between crocodile and elephant

7. What is the cause of the conflict of this story?
   a. The animals needed source of water
   b. The crocodile could not find water
   c. The elephant was not brave to fight against crocodile
   d. The crocodile was worried about running out of water

8. What happened to the elephant’s nose after it could escape from the crocodile?
   a. The elephant’s nose enlarges
   b. The elephant’s nose shortens
   c. The elephant’s nose becomes thinner
   d. The elephant’s nose elongates

9. …"The crocodile was irritated!..." (paragraph 4) the underlined word is similar in meaning to…
   a. Proud
   b. Hurtful
   c. Disappointed
   d. Angry

10. How did the writer describe the crocodile?
    a. Generous
    b. Selfish
    c. Boastful
    d. Brave

PART 2 (5 points)
You should spend about 10 minutes on this part
Read The following passage!

Carpenter Ants

Carpenter ants get their name because they build their nests in wood. This pest can cause significant damage to your house. There are many types of carpenter ants throughout the U.S. measuring in size from one-quarter inch (about the width of a pencil) for a worker carpenter ant to three-quarters of an inch (about the size of a quarter) for a queen carpenter ant.

Each colony is established by a single, fertilized queen. She starts her nest in a cavity in wood, where she raises her first brood of workers. She feeds them saliva and does not leave the nest or feed herself during this time.

When they are ready, those workers then get the job of gathering food to feed the next generation. Once mature, this first generation of worker ants work to increase the food supply for the colony. The colony population grows very rapidly. A colony can eventually produce 2,000 or more workers.

Complete the summary of the passage above using the list of words below! **(*Lengkapi ringkasan dari teks di atas dengan kata-kata yang tersedia di kotak bawah ini!)***

Fact about Carpenter ants

Carpenter is a given name of ants that (11)______ their nests in wood. This insect is (12)_____since it can damage house. Based on their size, queen ants are (13)______ than worker ants. Regarding their jobs, a queen ant raises its brood, while worker ants (14)_______food. The growth of carpenter ant colony is very (15)________. In fact, more than 2000 ant workers are generated by a colony.

collect create
bigger harmful
smaller feed slow
useful fast slower
break

PART 3 (5 points)

Read this brochure and answer the following questions in **NO MORE THAN 4 WORDS**. You should spend 15 minutes in this part. (*Bacalah brosur di bawah ini dan jawablah pertanyaan dengan tidak boleh lebih dari 4 kata. Anda disarankan mengerjakan bagian ini dalam waktu 15 menit*)

Bali Profound

5 days/4 nights

Travel like no other on this experience, created especially for nature Lovers and those seeking local encounters. Explore the rich history and culture with temple and local village visits, witness the breath-taking landscapes and nature highlights Bali has to offer.

Day 1-2: Bali Airport/Hotel – Sanur - Munduk

See one of the most famous temples in Bali, Tanah Lot, on the way to the UNESCO Heritage-listed Jatiluwih Rice Terraces.

Day 3: Munduk - Pemuteran

Start with a short trek through the rainforest, then jump on a traditional wooden boat as you cross the twin lakes of Buyan and Tamblingan.

Day 4: Pemuteran – Menjangan Island - Pemuteran

Enjoy a day trip to Menjangan Island for a chance to snorkel with turtles, sharks and plenty of other marine species.

Day 5: Pemuteran - Lovina

Rise early for a bird watching tour, or just simply relax for the day before heading to the Buddhist Monastery of Brahma Vihara.

Departures: Bali Airport/Hotel
Facilities: 4 nights' accommodation, breakfast daily, 4 lunches, 2 dinners, all entrance fees and tours, English speaking local guide and transportation by private air-conditioned vehicles
Price Guide From: $ 750

Questions

16. Who is probably reading this text?

17. What tourist resort in Bali is nominated as one of the UNESCO heritage?

18. In which day could you swim and watch animals under water?

19. What facility can be received by tourists who cannot speak Bahasa Indonesia or

Balinese?

20. What does the underlined word **local** refer to?

PART 4 (10 Points)

Read this story and answer the following questions in A SENTENCE. You should spend 15 minutes in this part. Remember you should paraphrase your answer. (*Bacalah cerita di bawah ini dan jawablah pertanyaan dalam* SATU KALIMAT. *Anda disarankan mengerjakan soal ini dalam waktu 15 menit dan gunakan paraphrase (kata-kata And sendiri) untuk menjawab soal tersebut*)

One day a monkey wanted to cross a river. He saw a crocodile in the river, so he asked the crocodile to take him across the other side. The crocodile told the monkey to jump on its back. Then the crocodile swam down the river.

Now, the crocodile was very hungry, so when it was in the middle of the river, it stopped and said to the monkey, "Monkey, my father is very sick. He must eat the heart of the monkey. Then he will be strong again." The monkey thought for a while. Then he told the crocodile to swim back to the river bank.

"What's for?" asked the crocodile. "Because I didn't bring my heart with me," said the monkey. "I left it under the tree, near some coconuts."

So, the crocodile turned around and swam back to the bank of the river. As soon as they reached the river bank, the monkey jumped off the crocodile's back and climbed up to the top of a tree. "Where is your heart?" asked the crocodile. "You are foolish," the monkey said to the crocodile. "Now I am free and you have nothing."

The monkey told the crocodile not to try to fool him again. The crocodile swam away, hungry.

Questions

21. How did the crocodile try to fool the monkey?
22. Why did the monkey ask the crocodile to swim back to the river?
23. What is the main idea of the paragraph 4 **(So, the crocodile…..)**?
24. From the passage, how will you describe the monkey? The monkey is smart
25. What are the moral values of the story?

References

(Cambridge handbooks for language teachers) Arthur Hughes-Testing for Language Teachers-Cambridge University Press (1989).pdf. (n.d.).

Azwandi (1996) referensi dari tesis.pdf. (n.d.).

Bachman, L. and A. P. (1996). *Language Testing in Practice*.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1984). *Taxonomy of educational objectives*.

Brown, H. D., & Abeywickrama, P. (2010). *Language assessment : principles and classroom practices*. Retrieved from http://webcat1.library.ubc.ca/vwebv/holdingsInfo?bibId=4182192

Harmer, J. (2004). *The practice of English language teaching*. New York: Longman.

Heaton, J. B. (John B. (1988). *Writing English language tests*. Retrieved from http://webcat2.library.ubc.ca/vwebv/holdingsInfo?bibId=647363

Heaton pp.88-134.pdf. (n.d.).

Hedges, T. (2000). *Teaching and Learning in the Language Classroom*.

Jeremy Harmer. (2003). The Practice of English Language Teaching. *ELT Journal*, Vol. 57, pp. 401–405. https://doi.org/10.1093/elt/57.4.401

Lie, A. (2007). EDUCATION POLICY AND EFL CURRICULUM IN INDONESIA: BETWEEN THE COMMITMENT TO COMPETENCE AND THE QUEST FOR HIGHER TEST SCORES Anita. *TEFLIN Journal*, *18*(1), 1–14. https://doi.org/http://www.teflin.org/journal/index.php/journal/article/view/130

McNamara, T. F. (Timothy F. (2000). *Language testing*. Retrieved from https://books.google.co.id/books/about/Language_Testing.html?id=RuxUkltYl_UC&redir_esc=y

Morrow, C. K. (2018). Communicative Language Testing. In *The TESOL Encyclopedia of English Language Teaching* (pp. 1–7). https://doi.org/10.1002/9781118784235.eelt0383

Nielsen, J. (2011). Cloze Test for Reading Comprehension. *Nielsen Norman Group: Jakob Nielson's Alertbox*. Retrieved from http://www.nngroup.com/articles/cloze-test-reading-comprehension/

Saukah, A., Cahyono, A. E., & Cahyono, A. E. (2015). NATIONAL EXAM IN INDONESIA AND ITS IMPLICATIONS TO THE TEACHING AND LEARNING OF ENGLISH. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *19*(2), 243–255. https://doi.org/10.21831/pep.v19i2.5583

Sulistyo, G. H. (2015). ENGLISH AS A MEASUREMENT STANDARD IN THE NATIONAL EXAMINATION: SOME GRASSROOTS' VOICE. *TEFLIN Journal*, *20*(1), 1–24. https://doi.org/10.15639/TEFLINJOURNAL.V20I1/1-24

Terry, R. M., & Hughes, A. (2006). Testing for Language Teachers. *The Modern Language Journal*, *74*(3), 383. https://doi.org/10.2307/327632

Urquhart, A. and C. W. (1998). *Reading in a Second Language: Process, Product and Practice.*