Available online at IJEE (Indonesian Journal of English Education) Website:

http://journal.uinjkt.ac.id/index.php/ijee

A MODEL OF AN ONLINE READING COMPREHENSION SUMMATIVE TEST FOR COLLEGE STUDENTS

Sofa[1], Gunadi H. Sulistyo[2]
[1]STKIP PGRI Jombang, Indonesia
[2]Universitas Negeri Malang, Indonesia
(sofataroki@gmail.com)

Received: 14th October 2017; Revised: 18th November 2017; Accepted: 24th December 2017

ABSTRACT

There is an emerging phenomenon in some universities including STKIP PGRI Jombang regarding a compelling need of a test that can replace the existing paper-and-pencil based reading comprehension test, which is conventional, impractical, and time consuming. To fulfill the need, a model of an online reading comprehension summative test was developed, involving a number of essential micro skills of reading. The design of the study was Educational Research and Development (R&D), involving 100 subjects in the try-out stage. The instruments used were interview guides and questionnaire. Based on the tryout analysis, the reliability was .779, in which thirty one items were categorized as valid items. For the ease of scoring and the balanced number of the indicators under interest, only 25 items were included in the model test. Based on the students' questionnaire, more than 80% subjects responded positively. The final product of this research was a set of an online reading comprehension test kit that includes the blueprint, the test (in form of paper and screenshot of the online version), the answer key, and the instruction to access the online test.

Key Words: online summative test; reading comprehension

ABSTRAK

Di beberapa universitas termasuk STKIP PGRI Jombang, muncul kebutuhan penting sebuah tes yang bisa menggantikan tes membaca berbasis paper-and-pencil sebelumnya yang konvensional, tidak praktis dan memakan banyak waktu. Untuk memenuhi kebutuhan tes yang bisa mengatasi masalah tersebut, dikembangkanlah sebuah model tes membaca sumatif online. Desain penelitian ini adalah penelitian pengembangan, yang melibatkan 100 subjek dalam tahap try-out. Instrumen yang digunakan adalah interview guide dan kuesioner. Berdasarkan analisis butir soal, nilai alpha atau reliabilititas adalah 0.779. 31 butir soal dikategorikan sebagai butir soal yang valid. Untuk kemudahan penilaian dan keseimbangan jumlah indikator yang diinginkan, hanya 25 butir soal yang digunakan dalam model tes. Berdasarkan kuesioner mahasiswa, lebih dari 80% subjek merespon secara positif. Produk akhir dari penelitian ini adalah satu set online reading comprehension test yang meliputi kisi-kisi, tes (dalam bentuk kertas dan screenshot versi online), kunci jawaban dan instruksi untuk mengakses tes online.

Kata Kunci: *tes sumatif online;* reading comprehension

How to Cite: Sofa., Sulistyo, G.H. (2017). A Model of an Online Reading Comprehension Summative Test for College Students. *IJEE (Indonesian Journal of English Education), 4*(2), 168-187. doi:10.15408/ijee.v4i2.8344

IJEE (Indonesian Journal of English Education), 4 (2), 2017, 165-184

P-ISSN: 2356-1777, E-ISSN: 2443-0390 | DOI: http://dx.doi.org/10.15408/ijee.v4i2.8344.

INTRODUCTION

Reading is one's inevitable daily needs. Sulistyo (2011, p.20) states that on one occasion, we read for information; on the other for enjoyment. This implies that reading comprehension plays a critical role in our daily lives. To reading teachers who are concerned with students' competence to read for information or knowledge through reading activities, there is a compelling need for them to always find an appropriate way to teach their students and to assess their reading comprehension with a greater attention as the ability to read is an important asset one must have on any occasion, let alone, in the digital era. Reading (critically) is believing; it is the window through which abundance of information is accessed.

A test is a subset of assessment (Brown, 2004, p.4). Further Brown (2004, p.4) states that a test is prepared administrative procedures that occur at identifiable times in a curriculum when learners muster all their faculties to offer peak performance, knowing that their responses are being measured and evaluated. In this way, learners are required to demonstrate their optimum competences elicited through tests in the form of manifest language behaviors.

To develop a good test, there are several criteria that need to be not only known but also fulfilled satisfactorily as a test is a set of data collection instruments that should function properly if accurate information about the learners is to be observed optimally to avoid the so-called gi-go effects – garbage in garbage out impacts. The first is validity. Gronlund and Linn (1990, p.47) state that validity refers to the appropriateness of the interpretations made from test courses and other evaluation results, with record to a particular use. It means that the result of the test should be meaningful, appropriate, informative, and useful. The second is reliability. Brown (2004, p.20) states that a reliable test is consistent and dependable in terms of the scores yielded by the testing procedures. If we give the same test to the same students on two different occasions, the test should yield about similar results. The third is practicality. Djiwandono (1996) states that practicality means something to do with the test administration, scoring, interpreting of the test results, even with the financial factors of the test administrations. Practicality may be concerned with economy in terms of resources, time, and energy. In line with the idea of Djiwandono (1996), Gronlund and Linn (1990) emphasize that there are some considerations that can be used to see the practicality of the test. The first is the use of test administration. For this purpose, the direction should be simple and clear, the subtest should be relatively few, and the timing of the test should not be too long. The second consideration is timing required for administration; it deals with allocated time to do the test. The other consideration is the ease of scoring which includes the clarity in the directions for scoring and simplicity in the scoring key. The following consideration is cost of testing which is important in selecting a test. The last is economy. Gronlund and Linn (1990, p.103) explain that testing should be relatively inexpensive and cost should not be a major consideration.

One of the types of tests that a teacher almost certainly needs to make is an achievement test. There are two types of achievement test: they are formative and summative tests (Brown, 2004, p. 48). A formative test aims at measuring the extent to which students have mastered the learning outcomes of a rather limited segment or instruction, such as a unit or a textbook chapter (Gronlund & Waugh, 2009, p.7). A summative test or it is also known as summative assessment aims to measure, or to summarize what students have grasped, and typically occurs at the end of a course or unit of instruction (Brown, 2004, p. 6). Popularly, the test that is mostly and continually carried out by classroom teacher is a summative test to know the students' mastery of the course. So, as it is crucial to know what the students have grasped, the concern about the summative test in reading needs to get greater attention.

Nowadays, considerable attention is paid to the nature a test as a part of three partite functions of assessment: assessment of learning, for learning, and that as learning. Earl, Katz, and WNCP team (2006, p. 55) state that assessment of learning refers to strategies designed to confirm what students know, demonstrate whether or not they have met curriculum outcomes or the goals of their individualized programs, or to certify proficiency and make decisions about students' future programs or placements. It is designed to provide evidence of achievement to parents, other educators, the students themselves, and sometimes to outside groups (e.g., employers, other educational institutions). It means that assessment is a crucial tool to show the students' learning mastery of the lesson based on the curriculum applied and further to decide what fits them in the future. Assessment of learning is in other words on the students' side. On the other hand, Earl, Katz, and WNCP team (2006, p. 29) also state that assessment for learning occurs

http://journal.uinjkt.ac.id/index.php/ijee | DOI: http://dx.doi.org/10.15408/ijee.v4i2.8344
P-ISSN: 2356-1777, E-ISSN: 2443-0390 | This is an open access article under CC-BY-SA license

throughout the learning process. It is designed to make each student's understanding visible so that teachers can decide what they can do to help students progress. In this part, teachers should investigate the students in the way they are studying, their problems, etc. to later find out the way to solve them and help them to understand the lesson. Assessment of learning is in other words on the teachers' side. The last is assessment as learning. Earl, Katz, and WNCP team (2006, p. 41) have stated that assessment as learning focusses on students and emphasizes assessment as a process of metacognition (knowledge of one's own thought processes) for students. It means that in the process of learning with their own understanding, students can do self-assessment to make sense of the information and use it for new learning under the guidance and the direction of the teacher. Assessment as learning in other words involves both the teachers' and students' side as well.

Supporting the ideas above, further Sulistyo (2015, p.5) states that assessment then implies an ongoing monitoring process on students' learning applied as soon as the teaching learning process begins, continuing up to the end of each class session. It informs teachers about their teaching effectiveness, students' learning progress, and even feedback on the level of implementation of a curriculum. As such, assessment is inseparably aligned to instruction. Further he also states that in a way, if carefully planned and implemented accurately, assessment can provide teachers with a source of useful information to reflect their teaching practices. It means that teaching cannot be separated from testing; they are linked to each other. Test results provide an important basis for the teacher to better design their teaching so that the teaching delivery can boost the students' performance in learning.

In recent days, reading from computer screens is becoming more and more common in human daily life as the amount of reading material available from online is rapidly increasing. This phenomenon has been seen in the field of language assessment such as computer-based tests (CBTs), computer-adaptive tests (CATs) and also TOEFL. As stated by Sulistyo (2009), for instance the advances in computing technology also boosts the presence of the new version of TOEFL, the iBT in 2005 which has been a significant shift from older TOEFL versions of computer based TOEFL (CBT for short) as well as paper-and-pencil based TOEFL (PBT, henceforth). This iBT version, as its name indicates, makes the functional use of information and communication technology (ICT). It means that the Internet in testing is already in broad use and it can support and optimize the assessment. One of the proofs that it is in fact quite important is that the growing demands of the services or software in online testing which increases year to year. Mason (1998) and Weisburgh (2003) (as cited in Hricko & Howell, 2006, p. 4) said, "The availability of assessment software to address these tasks is leading to assessment services becoming one of the fastest growing software niches, both in the corporate and in the educational markets.". Regardless the rapid growth of the demand in this area, development and implementation of this new mode of testing is currently in its initial stages. Therefore, sufficient empirical data, which would allow researchers to look into the soundness of computerized language tests with regard to construct validity and fairness, are yet to be available.

STKIP PGRI Jombang is one private university in operation in Jombang, East Java. In this university, the rapid use of the Internet network is also increasing but not yet functioned in the best way. Online assessment is in fact very helpful to not only students but also the lecturers to be the media in assessing processes. As Pallof and Pratt (2009, p. 3) put it to say, "The convenience of working online has proven to be very attractive to students and instructors alike." Further, Lynch (1997) (as cited in Millsap, 2000, p. 4) found that subjects responded more honestly on computer-administered tests than on paper and that the test-retest reliability was comparable for both groups. This means that online assessment offers convenience more than the traditional one in the now era.

In this university, in the Reading Comprehension 2 class, a substantial problem emerges. The test of the course is held by using a face-to-face interview to make the students explore more, to minimize the cheating, and to simplify the test. This face-to-face test is time consuming since with total students of forty has spent six hundreds minutes (10 hours) to assess student reading comprehension. A more efficient yet accurate and reliable test is then needed. The choice is an ICT-based test. By using an online test, the teacher can manage the time in the computer and score student reading performance in the test more quickly. In addition, online assessment is cost effective as lecturers do not need to copy the paper test to the whole students. As it has been said by Dowsing, Long, & Craven, (2000), Weisburgh, (2003) (as cited by Hricko & Howell, 2006, p.11) that "it has been proposed that one of the main advantages of using assessment software over manually assessing

http://journal.uinjkt.ac.id/index.php/ijee | DOI: http://dx.doi.org/10.15408/ijee.v4i2.8344
P-ISSN: 2356-1777, E-ISSN: 2443-0390 | This is an open access article under CC-BY-SA license

performance is primarily the savings in cost and time". In addition, computer-administered testing benefits include rapid up-dates, random item selection, test item banks, and automatic data collection and scoring (Millsap, 2000, p. 6). Practicality will also improve since the manual scoring will not be carried out by the lecturer like paper and pencil tests. As Weisburgh (2003) (cited in Hricko & Howell, 2006, p.11) said "Scoring and evaluating tests used to take a lot of manual effort, whereas software can dramatically reduce, or even eliminate, the manual effort, and results can be instantaneous". By all the facts elaborated above, this online test has huge possibility to be lower in cost. Another weakness point to be discussed is about the existing reading comprehension test is that the questions are in the form of oral questions, which implies impracticality of administration. Furthermore, these questions do not completely represent the indicators in the syllabus as the questions are only about the content, the generic structure and feature of the test and text building. The test only covers one type of text while the students must know all genres. This fact may lead to invalidity i.e. inaccuracy and error test results because of the teacher's subjectivity or tiredness. By having an online test, the problems will be solved as Krug (1989) reported that in an estimated ten percent of hand-scored objective tests, errors of one point or more in the final score were made. Computerized test administration ensures accurate test scores (as cited in Millsap, 2000, p.16).

Studies on the use of technology in testing have been conducted. A study by Sawaki (2001) aimed to examine the comparability of conventional and computerized tests of reading in a second language. The study used a survey design by a large sample as the subjects of the research. The general trends found in this study indicated that comprehension of computer-presented texts is, at best, as good as that of printed texts (Sawaki, 2001, p. 49). The second study was conducted by Noyes and Garland (2008) that investigated whether computer and paper-based tasks are equivalent. A survey design was conducted by reviewing literature and research. In the study, it is indicated that in some cases, paper and computerized tests were equivalent, but in some cases they were not for example in the form of the test. In addition to this finding, achievement of equivalence in computer-based and paper-based tasks poses a difficult problem. It is probably influenced by the test takers' confidence in using the computer, and other psychological factors.

Both studies basically state that computerized and paper-based tests cannot be said equivalent, but now in the year of 2017, it is very possible if they are equivalent or even computerized test will be more effective as people can see some of schools have conducted the computerized test (and the online one). Even now in senior high schools, the national examination is held online, too (*UNBK or Ujian Nasional Berbasis Komputer*). Teachers certification as well as lecturers certification is also conducted online. It means that online testing is broader in use, becoming more popular and offers more benefits despites its technical challenges.

Based on the context described in the previous section, the problem to be addressed in the present study is how can a model of an online reading comprehension test be conceptually and empirically developed to replace the existing test. The present study is therefore an attempt made to conceptually develop a set of an online reading comprehension test and empirically validate the reading comprehension test. Furthermore, the developed product is significant to replace the previous time consuming and non-effective test, to get the students' achievement score in the end of the lesson and to be a model for test developers (and/or lecturers) to develop a similar test for other reading courses (Reading Comprehension 1, Reading Comprehension 3, and Extensive Reading) or also other courses in general.

METHODS

The design of the test development model was adapted from Sulistyo (2015, p. 106). To meet the need of the present R&D research, some adaptations were carried out, so the model of the online test development used the following stages: conducting needs assessment, creating content specification/blueprint, blueprint expert review, prototype writing, prototype review, test installing, test and ICT expert review, try-out, item analysis, final form/publishing the final form.

The test installing or on-lining the test was carried out after the website namely www.sotaki.com was ready. The stages were *logging-in* as the admin to start the creating of the online test, *creating the course* to name the course which is Reading Comprehension 2 course, *creating the test* to name the test which is Summative test 1 and 2, *creating questions* to provide the questions, type of questions, the options, the texts in form of images, key answers and the score, *test setting* which includes timing and score viewing,

http://journal.uinjkt.ac.id/index.php/ijee | DOI: http://dx.doi.org/10.15408/ijee.v4i2.8344
P-ISSN: 2356-1777, E-ISSN: 2443-0390 | This is an open access article under CC-BY-SA license

users adding to input the user of the test in the database, *publishing* to bring the test online so it can be accessed by the students enrolled the course, the last is *result exporting* to take the data easily for later use. Data in this case refers to the students' names, scores, duration, timing and others in the excel format for later use in the item analysis stage. The name of the computer program utilized was Chamilo version 1.9.10.2.

The design of the needs assessment was qualitative. The instrument was interview to one Reading Comprehension 2 lecturer. It was about how the lecturer previously conducted the test, the form of the test, the reason why choosing certain form of test, the material included in the test and the availability of later online reading test for the students. After the information was gathered, the activity of collecting and preparing appropriate passages in various genres for the material in the body of the test started.

Three test and three ICT experts were invited to review and conceptually validate the products. The instrument used was in the form of questionnaire. In the test expert review, it was focusing on the items, the instruction (wording), and the construction of language test. The analysis was qualitatively carried out since the date got was in form of description. In the ICT expert review, it was focusing on the easiness of the instruction, the loading of the questions, the ease of the navigation menu, the readiness of the font and the *User Interface* generally.

The subjects of the tryout involved were 100 students of STKIP PGRI Jombang who had finished their Reading Comprehension 2 course. The decision of choosing the subjects employed simple random sampling. Latief (2012, p. 183) states that simple random sampling technique is the best technique in assuring the representativeness of the sample from the accessible population. It fits the needs of the samples since all students have an equal chance to be the representativeness of the sample. The try-out was carried out within two sessions to minimize the subjects to get tired.

A set of questionnaires is also addressed to the subjects. It is about the ease of the instruction, the ease of the questions, the time allotment, the suitability of the test and the material given in the class, the easiness of the texts, the length of the texts, the number of items and the level of difficulty of the items.

After conducting the informal try-out, the process of analyzing the test's result by using software called

ITEMAN 3.00 was carried out. The reliability is shown by the *alpha* score, which ranges from 1.00 for perfect reliability to 0.00 for completely unreliable (Ary et al., 2002, p. 261). The item validity can be known by the point-biserial correlation coefficient or symbolized by *r-pbis* coefficient. It is a statistic used to estimate the degree of relationship between naturally occurring dichotomous nominal scale and an interval or ratio scale (Brown, 2001, p.13), if the coefficient is > .2 it is categorized that the item is good.

Item difficulty is shown by the *proper correct* score (category easy range >.7, moderate range between .3-.7, and difficult is < .3) (Brown, 2001), item discrimination is presented in *p-bis* coefficients. The categorization of the item discrimination is shown below.

Table 1. Item Discrimination Categorization

Index range	Interpretation
≥ .40	Very good
.30-.39	Good
.20-.29	Fair
≤ .19	Poor

(Adapted from Djiwandono, 2011, p.230)

The effectiveness of distractor is important to be known as Brown (2004, p. 60) notes that the efficiency of distractor is the extent to which (a) the distracters "lure" a sufficient number of test takers, especially lower ability ones and (b) those responses are somewhat evenly distributed across all distractors. The efficiency of distractor can be known by the positive of negative value in *p-bis* key in each item. If there is a positive score of the efficiency distracter it means the distracter should be reviewed or changed.

FINDINGS AND DISCUSSIONS

Findings

The results of the development have been known after the research was carried out in STKIP PGRI Jombang.

The Result of Needs Assessment

It was found that the previous test was not practical, time consuming, and the material was only few than what it should be tested. The other fact from the interview was the availability of an online test in recent days has become a trend so that the availability of a model of a Reading Comprehension 2 summative test is needed to be carried out.

The Test Characteristics

Based on the syllabus of Reading Comprehension 2 course, the course intends to measure several micro reading skills that follow: identifying topics, identifying main ideas, identifying specific and detailed information (explicit and implicit), understanding the organization of ideas

http://journal.uinjkt.ac.id/index.php/ijee | DOI: http://dx.doi.org/10.15408/ijee.v4i2.8344

P-ISSN: 2356-1777, E-ISSN: 2443-0390 | This is an open access article under CC-BY-SA license

in texts, identifying reference, identifying vocabulary to derive meaning, identifying writer's tone or purpose and evaluating expressions in context. Based on the indicators stated in the syllabus then the item indicators can be used as a basis to develop test items. Sulistyo (2008) distinguished three domain of skills in reading, they are word attack, sentence attack and text attack skills. Based on the syllabus of Reading Comprehension 2, all these three skills are included. The level chosen is advocating the ideas by Crawley and Mountain (1995, p. 104-105) as follows: literal and inferential. The critical level is not included since the level of the students is intermediate and the critical level will be beyond of the scope of the competences for them. In the test, the literal level has 40% out of 100 items since it easier, inferential level have 60% out of 100 items. This percentage is taken for the inferential level dealing with inferring implicit information from the text which is more difficult but fit to the students' level. So, based on the percentage, there are 40 items in the literal level, and 60 items in the inferential level.

In this present study the passage theme is mostly those dealing with education, literature, science, life, and entertainment**.** They range from 212-495 words since the average students are still in the low level of intermediate. Although the biggest number is 495 words but the passage is in the level of 8th which means it is still standard in terms of the level.

The readability of the texts that were used is calculated by using Flesch-Kincaid Formula. The result can be seen in Table 2.

Table 2 The Result of Flesch-Kincaid Reading Ease Scores and Its Interpretation

No	The Genre of the Text	Flesch-Kincaid Reading Ease Score	Estimated Reading Grade	Interpretation/ Description Style
1	Narrative (The Necessity of Salt)	73.6	8th	Standard
2	Recount (Edgar Allan Poe)	54.5	High School Students	Fairly difficult
3	Spoof (Goat jumping into deep hole)	97.7	5th	Very easy
4	New Item (Tectonic earthquake sparked, Mt. Merapi's recent activity)	51.1	High School Students	Fairly difficult
5	Descriptive (Macquarie University)	44.5	College Students	Difficult
6	Report (A Museum)	41.4	College Students	Difficult
7	Explanation (How Was the Earth	46.1	College	Difficult

No	The Genre of the Text	Flesch-Kincaid Reading Ease Score	Estimated Reading Grade	Interpretation/ Description Style
	Formed?)		Students	
8	Procedure (How to make Candles)	81	6th	Easy
9	Analytical (Opportunity in the Global Financial Crisis)	39	College Students	Difficult
10	Hortatory (Should not Bring Mobile Phone to School)	67.1	8th	Standard
11	Discussion (The advantages and Disadvantages of Distance Learning)	48.3	College Students	Difficult
12	Review (2012 film)	56.2	High School Students	Fairly difficult
13	News Item (Strait of Malacca still not safe from pirates)	57.2	High School Students	Fairly difficult
14	Hortatory (Why Should Wearing a Helmet when Motorcycling)	56.9	High School Students	Fairly difficult
15	Analytical (Death Penalty)	68.1	8th	Standard
16	Explanation (How does body react to the heat?)	69	8th	Standard
17	Discussion (Pro and con of Computers for Students)	65.1	8th	Standard
18	Review (Twilight)	53.6	High School Students	Fairly difficult
19	Narrative (The Colossal UFO)	79.3	7th	Fairly easy
20	Recount	88	6th	Easy
21	Report (Dolphin)	56.8	High School Students	Fairly difficult

The Result of Expert Review

There were two domains of experts in the validation stage. There were test experts and the ICT experts. The test experts did validation twice, the first one was about the blueprint review validation and the second one was the online test or the product itself.

Blueprint Review

Based on the feedback from the three experts, the inputs were about the level of skills, the numbering of the items, the grammar, the order of the item indicators, and title for the texts and record for number of sub competences to be rationally balanced.

Test Review

The inputs were the running of the try-out which should be divided into two sessions to diminish tiredness of subjects which can influence the result, the readability, the order of questions based on paragraph, and the language mistakes. The last was about the sources, quality of options and

http://journal.uinjkt.ac.id/index.php/ijee | DOI: http://dx.doi.org/10.15408/ijee.v4i2.8344

P-ISSN: 2356-1777, E-ISSN: 2443-0390 |

grammar, the level of difficulty and face validity checking.

Suggestions from the three ICT experts were about the type of passage format, the attractiveness of the test, the use of auto-save for the saving, and the interface.

The Result of Item Analysis

Based on the ITEMAN analysis, it was found that the alpha reliability of the test was .653, which was categorized as acceptable and fair. The next analysis was item difficulty, which result is shown in the Table 3.

Table 3. The Results of Item Difficulty Analysis

Index Range	Category	Item Number	%
> 0.7	Easy	9,11,18,29,34,38,45,46,51,52, 56,57,60,69,72,87,88,90,94,97	20
0.3-0.7	Moderate	2,4,5,8,10,13,15,16,17,20,21,23,25,26,27,30,31,32,33,35,36,37,40,41,44,48,50, 54,55,61,62,63,65,66,68,71,73,74,75,76,77,78,79,80,81,82,83, 93,95,100	50
< 0.3	Difficult	1,3,6,7,12,14,19,22,24,28,39, 42,43,47,49,53,58,59,64,67,70, 84,85,86,89,91,92,96,98,99	30

Table 4. The Results of Item Discrimination Analysis

Index Range	Interpretation	Item Number	%
≥ .40	Very good	9,16,18,27,30,32,36,38,44,45,46, 64,70,74,78,83,85,90,92,94,95,97	22
.30-.39	Good	8,11,15,24,35,41,49,50,53,62, 63,65,66,72,75,87,100	17
.20-.29	Fair	1,4,6,12,13,23,26,29,33,40,48, 55,60,68,73,80,84,89,99	20
≤ .19	Poor	2,3,5,7,10,14,17,19,20,21,22,25,28, 31,34,37,39,42,43,47,51,52,54,56, 57,58,61,67,69,71,76,77,79, 81,82,86,88,91,93,96,98	41

Based on the result shown in the table 3, there are 20 easy items, 50 moderate items, and 30 difficult items.

In order to know how good the item in discriminating the low and high ability students, the analysis of item discrimination was carried out. From the ITEMAN version 3.00, the result is presented in the Table 4.

There are 22 items categorized as very good items, 17 items as good items, 20 items as fair items and 41 items as poor items.

Regarding the item validity, based on the result in the ITEMAN, the item validity is shown in Table 5.

From the result shown in the table 5, it can be seen that there are 31 items categorized as valid items and 69 items categorized as not valid items. These 69 items were dropped from the product and only 31 valid items were used.

The last analysis was the effectiveness of distractor. Based on the data from ITEMAN result analysis, there are 32 items which have suggested answer keys. These 32 items were dropped from the products and they were items numbers 2, 7, 19, 20, 21, 22, 24, 25, 28, 31, 39, 42, 43, 47, 51, 54, 56, 57, 58, 60, 61, 67, 71, 73, 77, 79, 81, 86, 91, 93, 98, 99.

The 31 good items were run to the ITEMAN 3.00 to be re-analyzed. The reliability is shown by the alpha score, which score is 0.779 and it can be categorized as good and can be used as the items in the test. The next thing is item difficulty, as shown in Table 6.

There are 8 items categorized as easy items, 17 items as moderate items and 6 items are difficult items.

Table 5 The Result of Item Validity Analysis

Index Range	**Item Number**	%	**Interpretation**
r > 0.2	8,9,15,16,18,27,30,35,36,38,44,45,46,49,53,62,64,65,66,70,74,78,83,85,90,92,94,95,97,100	31	Valid items
r < 0.2	1,2,3,4,5,6,7,10,11,12,13,14,17,19,20,21,22,23,24,25,26,28,29,31,32,33,34,37,39,40,41,42,43,47,48,50,51,52,54,55,56,57,58,59,60,61,63,67,68,69,71,72,73,75,76,77,79,80,81,82,84,86,86 88,89,91,93,96,98,99	69	Not valid items

Table 6 The Result of Item Difficulty Analysis

Index Range	**Category**	**Item Number**	**f**	%
> 0.7	Easy	2,5,11,13,14,26,28,30	8	26
0.3-0.7	Moderate	1,3,4,6,7,8,9,10,12,17,19,20,22,23,24,29,30	17	55
< 0.3	Difficult	15,16,18,21,25,27	6	19

From item difficulty, then the item discrimination has also run, the result has been shown in the Table 7.

Based on the result, 25 items are categorized as very good and 6 items are categorized as good which means that they can discriminate the students well.

Related to the item validity, all the 31 items are categorized as valid items and later for the easiness of scoring and the balanced number of the indicators under interest, the used items are only 25 items.

The Result of Students' Questionnaire Analysis

To gain the information about how the online test worked for the subjects' point of view, questionnaires with 10 multiple choice items and 2 essay questions were distributed to the 100 subjects. The result of the subjects' answer is presented in the table 8.

The typical format appearance of the product of the present study is presented in the figure 1.

Table 7. The Result of Item Discrimination Analysis

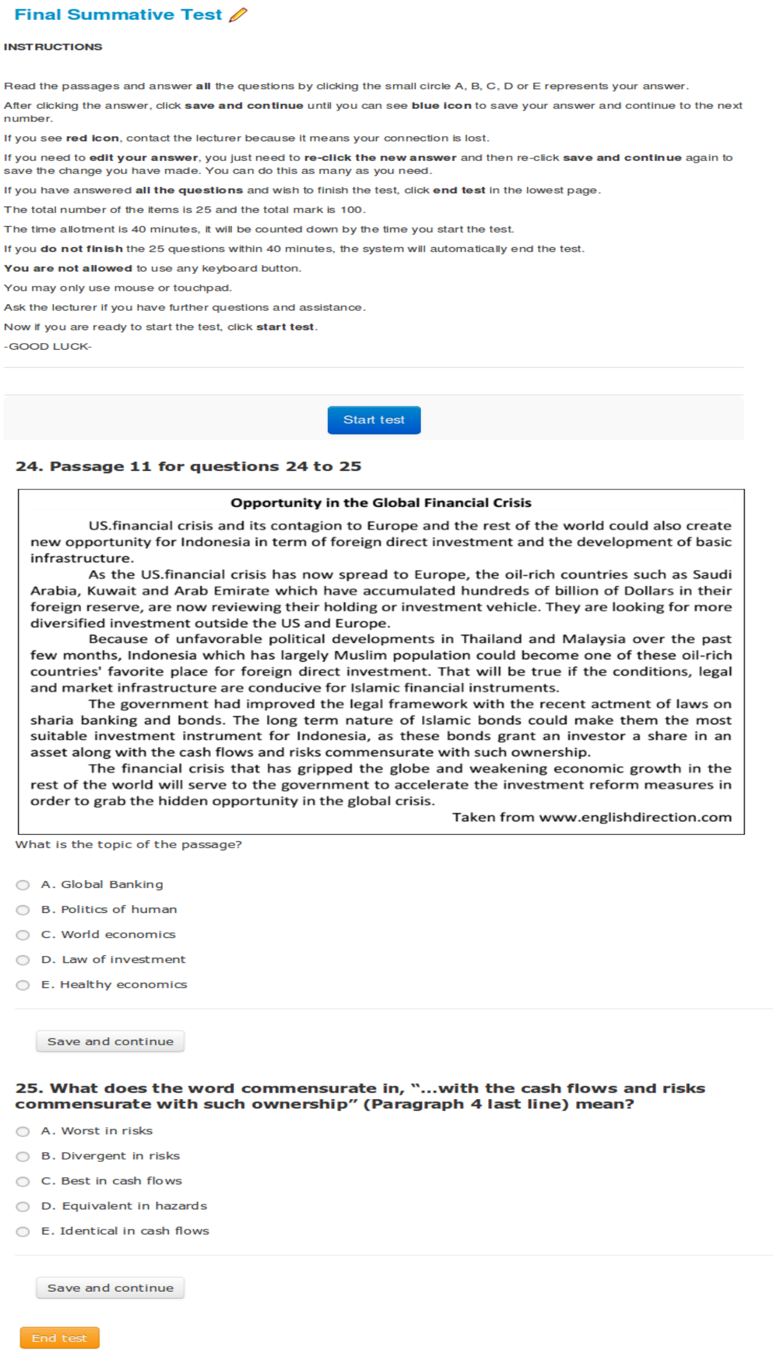
Index Range	Interpretation	Item Number	f	%
≥ .40	Very good	2,3,4,5,6,7,8,10,11,12,13,14,15, 17,18,21,22,23,24,25,26,27,28, 29,30	25	81
.30-.39	Good	1,9,16,19,20,31	6	19
.20-.29	Fair	-	-	
≤ .19	Poor	-	-	

Table 8. The Result of Item Validity Analysis

Index Range	Item Number	f	%	Interpretation
r > 0.2	1,2,3,4,5,6,7,8,9,10 11,12,13,14,15,16,17,18,9,20 21,22,23,24,25,26,27,28,29,30,31	31	100	Valid
r < 0.2	-	-		Not Valid

Table 9. The Result of Students' Questionnaire

No	Questions	Often		Seldom		Never		Sum
		f	%	f	%	f	%	
1	Before this test, how often have you been doing this type of online test?	5	5	46	46	49	49	100
		Very Easy		**Fairly Easy**		**Very Difficult**		
2	Are the instructions easy to be understood?	42	42	56	56	2	2	100
		Very Clear		**Fairly Clear**		**Less Clear**		
3	Is the way to answer the question clearly written?	65	65	33	33	1	1	99
		Very Easy		**Fairly Easy**		**Very Difficult**		
4	Generally, are the questions easy to be understood?	10	10	63	63	25	25	98
		Very Enough		**Fair**		**Less**		
5	Is the time allocation enough?	10	10	49	49	41	41	100
6	Generally, what do you think about the instructions to do the test?	97 % subjects said that the instructions are clear, simple and easy to be understood. 3 % said that the instruction is too many, but still it is clear.						
		Very Suitable		**Fairly Suitable**		**Less Suitable**		
7	Is the test suitable with the material given in the classroom?	30	30	64	64	6	6	100
		Very Easy		**Fairly Easy**		**Very Difficult**		
8	Based on the text difficulty level, are the texts easy to be understood?	5	5	50	50	44	44	99
		Too Many		**Fair**		**Less**		
9	Based on the number of the items, how are they?	34	34	65	65	1	1	100
		Too Long		**Fair**		**Too Short**		
10	Based on the lengths of the texts, how are they?	40	40	56	56	3	3	99
		Very Difficult		**Fairly Difficult**		**Easy**		
11	Based on the difficulty level generally, how are they?	16	16	81	81	3	3	100
12	Generally, what is your opinion about this online reading comprehension 2 test?	86 % subjects said that the online test is good, interesting, effective, practical, has less chance of cheating, fun and do not need to open the page too often, go along with the era, but 14 % subjects somehow said it also makes the eyes tired, the time is less and it is difficult.						



Final Summative Test

INSTRUCTIONS

Read the passages and answer **all** the questions by clicking the small circle A, B, C, D or E represents your answer.

After clicking the answer, click **save and continue** until you can see **blue icon** to save your answer and continue to the next number.

If you see **red icon**, contact the lecturer because it means your connection is lost.

If you need to **edit your answer**, you just need to **re-click the new answer** and then re-click **save and continue** again to save the change you have made. You can do this as many as you need.

If you have answered **all the questions** and wish to finish the test, click **end test** in the lowest page.

The total number of the items is 25 and the total mark is 100.

The time allotment is 40 minutes, it will be counted down by the time you start the test.

If you **do not finish** the 25 questions within 40 minutes, the system will automatically end the test.

You are not allowed to use any keyboard button.

You may only use mouse or touchpad.

Ask the lecturer if you have further questions and assistance.

Now if you are ready to start the test, click **start test**.

-GOOD LUCK-

24. Passage 11 for questions 24 to 25

Opportunity in the Global Financial Crisis

US.financial crisis and its contagion to Europe and the rest of the world could also create new opportunity for Indonesia in term of foreign direct investment and the development of basic infrastructure.

As the US.financial crisis has now spread to Europe, the oil-rich countries such as Saudi Arabia, Kuwait and Arab Emirate which have accumulated hundreds of billion of Dollars in their foreign reserve, are now reviewing their holding or investment vehicle. They are looking for more diversified investment outside the US and Europe.

Because of unfavorable political developments in Thailand and Malaysia over the past few months, Indonesia which has largely Muslim population could become one of these oil-rich countries' favorite place for foreign direct investment. That will be true if the conditions, legal and market infrastructure are conducive for Islamic financial instruments.

The government had improved the legal framework with the recent actment of laws on sharia banking and bonds. The long term nature of Islamic bonds could make them the most suitable investment instrument for Indonesia, as these bonds grant an investor a share in an asset along with the cash flows and risks commensurate with such ownership.

The financial crisis that has gripped the globe and weakening economic growth in the rest of the world will serve to the government to accelerate the investment reform measures in order to grab the hidden opportunity in the global crisis.

Taken from www.englishdirection.com

What is the topic of the passage?

○ A. Global Banking

○ B. Politics of human

○ C. World economics

○ D. Law of investment

○ E. Healthy economics

25. What does the word commensurate in, "...with the cash flows and risks commensurate with such ownership" (Paragraph 4 last line) mean?

○ A. Worst in risks

○ B. Divergent in risks

○ C. Best in cash flows

○ D. Equivalent in hazards

○ E. Identical in cash flows

Figure 1. Final Summative Test Online

Discussion

The result of needs assessment has revealed all the problems in the previous test, which is considered to be impractical. This online test is practical since it is easy in administration, easy in scoring and interpreting the result. The previous test is time consuming while this online test is time effective. The previous test covers only one genre while this online test covers all of the genres. The additional benefits of this online test are that this online test is cost effective and up to date. All the result of the needs assessment indicated that the online test has fulfilled the theory of criteria of a good test elaborated above by Djiwandono (1996) and Gronlund & Linn (1990). This online test is also has the advantages as what previous study by Noyes and Garland (2008) elaborated for example the richness of interface, accessible at home, less error in administration, online scoring which is greater in accuracy and less human error, and cost saving. Singh, Rylander & Mims (2012) also support the increase use of the Internet. They said that as preferences for online learning increases, mostly due to the convenience and flexibility it offers students, universities find themselves increasing the number of online format courses to meet the growing demand (p.96). Coiro (2014, p.12) added that there are many opportunities when students do learning activities online, such as question, wonder, and think more deeply about things with puzzle games, creating digital products , it also offers time for students to practice questioning, locating, evaluating, and synthesizing information collaboratively with a partner or in a small group (Coiro, p.16). Based on those facts, it is argued that the test in the present study can overcome technical problems in the previous tests ever developed.

As the items analysis was run, most of the items, the 69 items were invalid items, which should be dropped from the test. This means that only 31 items can be saved and used for the test. The reliability that is shown by the Alpha coefficient is .779, which can be categorized as good. The coefficient demonstrated that this scores generated from the test are consistent and reliable across measurement to show the real student's performance. The result indicates that this online test has one more quality of a good test in terms of reliability as explained above by Brown (2004).

The questionnaires show that most students respond positively toward the online test. Most of them respond that the test instruction (the instruction to operate the test and to answer the question) is generally easily understood which means the instruction is clear, causing no bias. They also respond positively that the questions are easily understood. The time is sufficient which means that the texts, the questions, and the time allocation are proportional to their level. The result is in line with what stated above by Gronlund and Linn (1990) about practicality and Zandvliet and Farragher, (1997) as cited in Noyes and Garland (2008, p.1369) about the advantages of computer testing. The material used in the test are suitable which means the test does not cover the material that was never taught in the classroom. A number of the subjects (44%) stated that the passages are difficult, which possibly because some students are actually in the lower proficiency level while this test is designed for the intermediate ones as it is stated in the syllabus. This fact also could be a reason behind the non-optimal alpha score. The last is about the subjects' opinion. Although few subjects say that the online test makes the eyes tired, mostly they say that the online test is good, interesting, effective, fun, practical, minimizing the chance of cheating. They also think that they do not need to open the page too often, and the test goes along with the ICT era. This means that the availability of this online test overcomes the problem emanating from the previous test used.

CONCLUSION AND SUGGESTION

The conclusions comprise the strengths and also the weaknesses of the product of this research. Related to the strengths, first, the product of this research can be a model of an online reading summative test in STKIP PGRI Jombang. Second, based on the try-out stage, it is shown that some items of the proposed test are valid and reliable. The product of this research is packaged into one part. It covers the blueprint, the test in the paper printed form and the screenshot of the online version, the answer key, and the instruction for access to the online test. As the product has strengths, it also has weaknesses. The final product of this test only consists of 25 items due to the elimination of the non-valid items. The reading level is not in the precise percentage as this study suggested. This product has no construct validity process to reveal the psychological quality of the students. In addition, this study is still at the automaticity process from the paper-based format to the computer-format one.

Some suggestions are presented after completing the whole processes in conducting this research. This online test can be a model for other reading courses and also other courses in general in conducting tests since it has been validated. This product can be an insight for the effectiveness of an online reading test in enhancing students' reading motivation with better qualifications for example random setting. And as this research had limited subjects (only 100 subjects), it is suggested that future researcher can have larger subjects to gain more reliable and valid result. Further, although low, but as this test still open the chance for the students to do the cheating, so the researcher will be working on the online test in randomized options. This attempt is hoped to not only diminish the cheating action but also increase the students' independence and self-esteem.

REFERENCES

Ary, D., Lucy, CJ., & Asghar, R. (2002). *Introduction to Research in Education. (Sixt Edition)*. Belmont: Wadsworth.

Brown, H.D. (2004). *Language Assessment: Principles and Classroom Practices*. White Plains: Pearson Education.

Brown, H.D. (2007). *Teaching by Principles: An Interactive Approach to Language Pedagogy (3rded)*. White Plains, NY: Pearson Education.

Brown, J.D. (2001). Statistics corner: Questions and Answers about Language Testing Statistics: Point Biserial Correlation Coefficients. *Shiken*: *JLT Testing & Evlution SIG Newsletter*. 5 (3):13-17. Retrieved from http://jalt.org./test/bro_12/htm on 12/12/2013.

Coiro, J. (2014). Online Reading Comprehension: Challenges and Opportunities. Retrieved from https://www.researchgate.net/publication/277897021_online_reading_comprehension_challenges_and_opportunities on 5/9/2018.

Djiwandono, M.S. (1996). *Tes Bahasa Dalam Pengajaran*. Bandung: Penerbit ITB

Djiwandono, M.S. (2011). *Tes Bahasa: Pegangan bagi Pengajar Bahasa*. Jakarta: PT Indeks.

Earl, L, Steven, K., & WNCP team. (2006). *Rethinking Classroom Assessment with Purpose in Mind: Assessment for Learning, Assessment as Learning, Assessment of Learning*. Western and Northern Canada: Ministers of Education.

Gronlund, N. E., & Linn, R. L. (1990). *Measurement and Evaluation in Teaching (Sixth Edition)*. New York: Macmillan.

Gronlund, N.E., & C. Keith, W. (2009). *Assessment of Students Achievement. (Ninth Edition)*. Upper Saddle River: Pearson Education.

Hricko, M., & Scott L.H. (2006). *Online Assessment and Measurement: Foundations and Challenges*. Hershey: Information Science Publishing.

Latief, M.A. (2012). *Research Methods on Language Learning: An introduction*. Malang: UM Press.

Millsap, C.M. (2000). *Comparison of Computer Testing Versus Traditional Paper and Pencil Testing*. Published Dissertation. Denton: Department of

Philosophy University of North Texas.

Noyes, J.M. & Garland, K.J. (2008). Computer- vs. paper-based tasks: Are they equivalent?. *Ergonomics*. Vol. 51, No. 9 pp. 1352–1375.

Sawaki, Y. (2001). Comparability of Conventional and Computerized Tests of Reading in a Second Language. *Language Learning & Technology*. Vol. 5, No. 2 pp. 38-59.

Singh, S., Rylander, D.H., Mims, T.C. (2012). Efficiency of Online vs. Offline Learning: A Comparison of Inputs and Outcome. *International Journal of Business, Humanities and Technology. Vol. 2 No. 1; January* 2012. Retrieved from http://ijbhtnet.com/journals/Vol_2_No_1_January_2012/12.pdf on 5/9/2018.

Sulistyo, G.H. (2007). *Tests, Assessment, and Measurement in English as a Second Language at Schools*. Malang: State University of Malang Press.

Sulistyo, G.H. (2009). TOEFL in a Brief Historical Overview from PBT to IBT. Retrieved from http://sastra.um.ac.id/ on 3/9/2012.

Sulistyo, G.H. (2011). *Reading for Meaning: Theories, Teaching Strategies, and Assessment*. Malang: Pustaka Kaiswaran.

Sulistyo, G.H. (2015). *EFL Learning Assessment at Schools: An Introduction to Its Basic and Principles*. Malang: Bintang Sejahtera.

STKIP PGRI Jombang. (2010). *Syllabus of Reading2*. Jombang: English Department.