IRJE (Indonesian Research Journal in Education)
Published by Jambi University, the Graduate School, Doctoral Program in Education
E-ISSN: 2580-5711

Developing an Indonesian Reading Proficiency Test for BIPA Learners

ANDIKA EKO PRASETIYO[1]

Abstract

The use of Indonesian proficiency tests for non-native speakers of Bahasa Indonesia is still equated with tests for native speakers. This has become a point of debate for many teachers and experts of Indonesian for Foreigners (*Bahasa Indonesia untuk Penutur Asing - BIPA*). The crux of the debate focuses on whether the same proficiency test should be used for both native speakers (NS) and non-native speakers (NNS) alike, or whether separate tests should be developed. In accordance with the peculiarities of *Bahasa Indonesia*, Indonesian proficiency tests for NS and NNS should be differentiated. The underdevelopment of specialized proficiency tests for NNS can be explained by the fact that *Bahasa Indonesia* is not one of the dominant languages learned in the world today. This research aims to develop materials for an Indonesian proficiency test for NNS. The development of the test focuses on reading comprehension. To advance development of the test, discussions of the processes for defining the theoretical construct as well as empirical analysis of students' result were combined. The method used in this study involved expert review, text readability analysis, and item analysis. The findings show that the test items developed can be used to test students' proficiency, particularly in reading comprehension.

Keywords
BIPA, foreign language, Indonesian, language testing, reading

[1] A fulltime graduate student at the University of Melbourne, Melbourne, Australia; andikaekop@gmail.com

Introduction

A proficiency test developed for Indonesian language learners is called *Uji Kemahiran Berbahasa Indonesia (UKBI)* is used as an Indonesian language proficiency test for both foreign speakers and native speakers. However, the assessment instruments used to test native speakers (NS) and non-native speakers (NNS) should be differentiated since the test objectives and test-takers are distinct.

Based on this issue, we sought to conduct research into the development of Indonesian language tests that are used to measure the Indonesian reading ability of NNS. Therefore, the product resulting from this study is a proficiency test that was developed for Indonesian language learners. Furthermore, the test developed can form a recommendation and an alternative for the language center as a measurement tool in addition to *UKBI*.

The test focuses on the reading comprehension aspect of testing. This test material will refer to the CEFR curriculum in which, at the advanced level, speakers must be able to read and comprehend of all forms of written language including structurally and linguistically complex texts such as abstracts, manuals, scientific articles, and literary works. A pilot study has also been included to ensure that the developed test has reliability and readability. The test was then administered to the students at the University of Melbourne, Semester 2 2018, in the subject Indonesian 3.

The objectives of this study, three main questions will be explored. (1) Based on the content validity, does the test reflect the course objectives? (2) What is the level of difficulty, index of discrimination, and distracters of each item? (3) What revisions are to be made of test items based on the test analysis?

Literature Review

Reading comprehension

The skill of reading comprehension is one of the most critical aspects of learning a language. For this reason, reading tests are now a crucial part of most major foreign language assessment protocols including TOEFL, IELTS, and TOEIC. In the last decade, many studies have investigated reading comprehension tests for foreign languages (e.g., Bernhardt, 1983; Gorsuch & Taguchi, 2008; Gorsuch & Taguchi, 2010; Keenan, Betjemann, & Olson, 2008; Rahmiati & Emaliana, 2017; Taguchi, Gorsuch, Takayasu-Maass, & Snipp, 2012; Taguchi, Takayasu-Maass, & Gorsuch, 2004). Tests of reading comprehension have become some of the most important instruments with which to measure a learner's proficiency in a foreign language acquisition. This is because reading tests have demanding characteristics in terms of cognition, requiring the synchronisation of memory, attention, as well as comprehension (Sellers, 2000).

In addition, reading comprehension tests can also involve both low order and higher orders of thinking (Rahmiati & Emaliana, 2017). This can be seen from the variety of texts presented in reading comprehension tests, including expositions, news, and literature. Reading comprehension tests also require several key characteristics in order to be considered sound and reliable. Firstly, the test must have validity and a relevant construct (Hughes, 2003). Secondly, the items included in a reading test should have reliable and consistent characteristics in terms of producing results (Brown, 2004). Thirdly, the reading test should be able to distinguish the level attained by the learner, such as whether the learner has achieved a primary, intermediate, or advanced level of language proficiency (Heaton, 1989). Finally, in terms of practicality, reading tests should also be effective and efficient to administer (Weir, 1990).

Question types in reading test

There has also been some discussion about the types of the questions that should be included in such reading tests. In his study, Pyrczak (1975) found that there was no significant difference in results between students who read the passage before answering, and students who did not read the passage when completing a multiple choice reading test. In addition, Jones (1977) argues that a proper foreign language reading test should utilise model translation. However, he stated that it would be difficult to assess since it might be more focused on grammatical aspects rather than meaning.

Meanwhile, Cranney (1972) suggests that the method of cloze reading is an excellent way to test reading skills. He also said that cloze reading is easy to produce and to score. Shohamy (1981), however, found that students have a negative perspective towards cloze reading. She found that students often felt that cloze tests were tough and frustrating. On the other hand, there is a study which supports the use of the multiple-choice method in reading test. Gorjian (2013) argues that tests with large numbers of participants are more suitable to the multiple choice question type. Based on this final theory, we have chosen to use a multiple-choice type format in developing a reading comprehension test for Bahasa Indonesia as a foreign language.

Empirical studies on foreign language test in reading

Regarding published research on the development of foreign language tests, several studies have investigated the area of testing for reading comprehension (e.g., Nindyaningrum, 2018; Rahmiati & Emaliana, 2017; Saifudin, Suwandi, & Setiawan, 2014). However, studies specifically exploring reading comprehension of Bahasa Indonesia as a foreign language are limited. Saifudin et al., (2014) developed an instrument that can be used as a measure of the proficiency of NNS in *Bahasa Indonesia*. In the development of this instrument, they adopted the international standardized test model, IELTS. However, they focused on all of the skills of language acquisition and proficiency, not simply reading ability and comprehension. Rahmiati and Emaliana, (2017) also developed a reading comprehension test, but only for English as a foreign language. The development of the reading test in her study focuses on both higher and lower order thinking of the students.

Moreover, the type questions which were developed in her study were multiple choice format. Nindyaningrum (2018) conducted a study on the development of reading comprehension test instruments for NNS. The instrument that she developed can be used to measure the reading proficiency of Indonesian learners. This piece of research will mirror the study by Nindyaningrum (2018). It should be noted however, that Nindyaningrum (2018) did not perform a test item analysis including, for example, descriptive statistics, facilitation value, discrimination index, nor distractor analysis in her study. To try and address this issue, this study will therefore also develop a reading test analysis that tries to include such items.

Descriptive statistics, item analysis, item facility, item discrimination

Descriptive statistics in developing reading proficiency test are beneficial to examine the students' score distributions in the test. The aim of the proficiency test is to distinguish the level of learners' competencies in comprehending the reading. Therefore, the score distribution may indicate the level of the students' competencies, which are low, medium, and advanced. On the other hand, the score distribution can also indicate the level of the difficulty of the questions (Brown & Hudson, 2002). To examine the score distribution in the reading test, Brown and Hudson (2002) suggest using the measures of central tendency, i.e. mean, mode and median which is part of the descriptive statistics.

Item analysis includes item facility and item discrimination. These two types of analysis are used to determine which items of the questions can be chosen and which items of questions need to be changed. The level of difficulty or the measurement of whether the test item is easy or difficult can be identified by calculating the value of item facilities, also known as item difficulty.

In terms of measuring the item facility in proficiency reading tests, there are two methods for calculation. The first way to identify the item difficulty is by measuring the number of correct items answered by test takers and then divide by the total number of test-takers (Bachman, 2004; Farhady, 2012). In addition, an alternative method is proposed by Bachman (2004, p.122) who suggests calculating "the proportion of test takers who chose the different distractors" in order to measure the difficulty level of items.

Item discrimination in proficiency tests refers to the ability of the item to distinguish the level of test takers' proficiencies, such as that of basic, intermediate, and advanced learners. In order to determine the item discrimination value, the number of test takers who give the correct answer to each test item is calculated and these numbers are used in a formula for discrimination index (Bachman, 2004). The value range of item discrimination is between -1 and +1. A higher value of item discrimination is better. Higher item discrimination indicates that the item is very effective for identifying the proficiency level of test takers (Farhady, 2012).

Methodology

To strengthen the development of the test, the processes of theoretical construct definition are discussed along with empirical analysis of students' test results. The method used in this study involved several steps, such as the expert judgment, analysis of text readability, and analysis of test items. The outline of test specifications was designed before creating the test items.

There are several steps in developing the test: 1) developing outline of test specifications 2) writing the blue print of the test; 3) writing the test items; 4) validating the test by an expert; 5) administering the test; 6) analyzing the test result; and 7) revising the test. The test was developed to measure the comprehension of learners in the reading of different kinds of text genres.

Each item of the test relates to readings in *Bahasa Indonesia* of various types, such as exposition text, news, and literature in the form of short stories. Each text has a length of about 136 - 295 words adapted from various sources. Topics and the features of *Bahasa Indonesia* are carefully transcribed into text, questions, and multiple choice alternatives.

The micro skills tested include understanding topics, main ideas, supporting details, implied details, word meaning, as well as drawing conclusions from texts. Moreover, the expert consulted, a University of Melbourne lecturer, stated that the test developed is feasible and ready to be used for testing. Based on this evidence, we conclude that the content and item of the reading test is valid.

Participants

This study was conducted at the University of Melbourne involving 32 students between the ages of 18 and 27. Each subject was taken from one of either two different classes, but still in the same subject, Indonesian 3 which is a *Bahasa Indonesia* class considered to be at intermediate level. The students consist of 16 males and 16 females.

All were NNS of *Bahasa Indonesia* originating from 7 different countries, namely Australia (*N* = *24*), Malaysia (*N* = *1*), Brunei Darussalam (*N* = *1*), England (*N* = *1*), USA (*N* = *1*), Singapore (*N* = *1*), and Indonesia (*N* = *1*). It should be noted that the one student from Indonesia has lived for a long time in Australia and uses English as their everyday language. Furthermore, when asked to self-rate their level of proficiency, 3 students were rated as advanced learners, 19 as intermediate, and 10 below intermediate.

Regarding the duration of learning *Bahasa Indonesia*, 14 students had been studying the language for less than 1 year, 7 students studying about 2 - 5 years, and 11 students studying for 6 years and above. Also, in terms of the level of reading intensity in *Bahasa Indonesia*, for example through magazines, books, and newspapers, 31% of students stated that they never do such reading, 44% said rarely, and 25% said do some reading but not extensive.

Procedures and test item writing and piloting test

In terms of the procedures of the test development, there were two main methodologies that were utilized, which were test development (test item writing and piloting test) and test administration. I developed test items based on an example of a *Bahasa Indonesia* proficiency test instrument. However, they are designed to meet the purpose of the test which is to measure the proficiency of NNS of *Bahasa Indonesia* in reading. Furthermore, the questions are based on three different types of authentic text, which are exposition, news, and literature texts. Initially, we developed a test with a variety of topics and different types of texts, including a personal letter, news, and literature. The first text is a personal letter (constructed by the researcher).

The second text is news about a museum fire that occurred in Jakarta written by Nurito (2018). The last text is a literature text, which is a short story entitled "*Anak Kebanggan*" by Navis (2018). The story was edited to be of an appropriate length and readability. No other significant changes were made to each of the three texts, other than length and readability. Regarding the number of questions, there are 20 question, with 5, 7, 8 questions for texts 1, 2, and 3, respectively. Also, each item has one correct answer and 3 distractors. All the items were aimed to measure learners of *Bahasa Indonesia* in their reading comprehension.

Then, we designed 20 multiple-choice questions based on three short texts, a letter text, a news text, and a short story. The length of time allotted to do the test was 20 minutes. Furthermore, to measure test readability, we conducted a pilot test pilot with 3 NNS students to ensure that the test developed were feasible. In addition, we also consulted both via email and direct discussion with the lecturer of Indonesian 3. This was with a view to gaining more input and feedback regarding the test.

Based on the test pilot, the lecturer gave a positive response to the test. However, some of the questions in text 1 (letters) are too easy, and most students answered them correctly. In addition, the teacher also gave input in our discussion that the comprehension of the letter text did not match the construct of relevance to real life. Therefore, we revised the first text by transforming it into an exposition text. In addition, we also adjusted the layout by providing row numbers on the left side of the text.

Administration of the test

The test was conducted twice in two different sessions of the same subject, Indonesian 3. The tests were administered on 21 and 22 May 2018 with time duration of 45 minutes in each class. Before conducting the test, we were assisted by the teacher explaining to the students about the purpose of the research. Teachers also helped by explaining that the tests might help them to prepare for final exams or improve their proficiency in *Bahasa Indonesia*, especially in reading.

Participants were given 25 minutes to complete 20 reading questions. Before working on the questions, students filled out a list of background questionnaires, including names, personal information, country of origin, previous experiences while learning *Bahasa Indonesia*, and self-assessment. This data was collected in addition to test score results to help identify other variables that may contribute to the variability of the test results.

Findings

Descriptive statistics of the test results

Table 1 shows the reading comprehension level of Indonesian as a foreign language in this study had a mean of 11.7 out of 20 and (*SD* = 3.6). This means that 58% of the test items were able to be answered correctly by students. Furthermore, the results also show that the lowest value is 20% (*N=1*), while the highest score is 100% *(N=1)*. To determine the learners' level, we adopted the TOEFL level rubric, which describes elementary level (0% - 50%), low intermediate (51% - 75%) high intermediate (76% - 85%), advanced (86% - 100%).

Based on this scale, it can be reported that 11 students were at the elementary level, 17 students were at low intermediate, 1 student was at high intermediate level, and 3 were at advanced level. These results indicate that the test developed was appropriate and not too easy or too difficult for NNS of *Bahasa Indonesia*.

Table 1. *Descriptive statistics of the test*

ID	Score / 20	Level	ID	Score / 20	Level
1	12	Low Intermediate	17	13	Low Intermediate
2	10	Elementary	18	16	High Intermediate
3	9	Elementary	19	7	Elementary
4	6	Elementary	20	15	Low Intermediate
5	9	Elementary	21	14	Low Intermediate
6	7	Elementary	22	15	Low Intermediate
7	11	Low Intermediate	23	13	Low Intermediate
8	8	Elementary	24	13	Low Intermediate
9	11	Low Intermediate	25	20	Advanced
10	11	Low Intermediate	26	4	Elementary
11	13	Low Intermediate	27	14	Low Intermediate
12	13	Low Intermediate	28	11	Low Intermediate
13	14	Low Intermediate	29	18	Advanced
14	9	Elementary	30	11	Low Intermediate
15	7	Elementary	31	9	Elementary
16	13	Low Intermediate	32	18	Advanced

The difficulty level of each item in the test

The item difficulty for each item was analyzed by using MS Excel (IF and point-biserials.xlsx). Each item has a range of 0.00 to 1.00. The interpretation of that number is the higher the value, the easier the test item. Furthermore, based on Djiwandono (1996), the indicators of item difficulty are as follows, easy (0.7 – 1), moderate (0.3 – 0.7), difficult (0 – 0.3). Table 2 below presents the facility value of each item.

Table 2. *The facility value of each item*

Item No.	Item Facility Value	Interpretation
1	0.91	Easy
2	0.91	Easy
3	0.81	Easy
4	0.28	Difficult
5	0.69	Moderate
6	0.03	Too Difficult
7	0.78	Easy
8	0.66	Moderate
9	0.81	Easy
10	0.47	Moderate
11	0.88	Easy
12	1	Too Easy
13	0.63	Moderate
14	0.41	Moderate
15	0.88	Easy
16	0.47	Moderate
17	0.22	Difficult
18	0.19	Difficult
19	0.25	Difficult
20	0.44	Moderate

The data in table 2 show that the item facility value is varied. A total of 8 (40%) items are categorized as easy items, 7 (35%) as moderate, and 5 (25%) as difficult items. Based on this analysis, the distribution of the proportion of item difficulty, such as easy, moderate, and difficult was balanced and appropriate, because the test developed is a proficiency test. However, there is one item that was too difficult with IF 0.03 (item number 6), and one item was too easy with IF 1 (item number 12). Thus, these two items will be revised to make the facility value in each item more appropriate.

The discrimination index of each item in test

To see how well an item can differentiate between higher and lower level learners, the discrimination index values can be examined. The higher the discrimination index value, the better the item distinguishes between the higher and lower level test takers (Farhady, 2012). Furthermore, if the discrimination index value equals 0, it indicates that low and high learners show the same performance, whereas a negative discrimination index value of indicates that lower students perform better than the higher students. D value 0 and - suggests that the item needs to be deleted or revised. The analysis of discrimination index values in this study was conducted using SPSS.

Table 3. *The discrimination index value*

Item No.	DI Value	Interpretation
1	.161	Enough
2	.161	Enough
3	.376	Good
4	.205	Enough
5	.292	Enough
6	.371	Good
7	.267	Enough
8	.372	Good
9	.210	Enough
10	.459	Very good
11	.414	Very good
12	0	No discrimination
13	.217	Enough
14	.603	Very good
15	.220	Enough
16	.440	Very good
17	.458	Very good
18	.492	Very good
19	.569	Very good
20	.621	Very good

Based on table 3, most values indicate good items *($N = 11$, $D > 0.3$)*. However, a total of 8 test items indicate as an adequate test (0.11 - 0.29), while one item indicates the item should be revised because the item shows that the number of discrimination index value is 0 or no discrimination.

Item distracters

Another way to investigate the item difficulty is by calculating "the proportion of test takers who chose the different distractors" (Bachman, 2004, p.122). The performance of each distractor can be seen from the analysis of item distracters. The test item which has distracters that are never chosen are useless and need revision. However, distractors that attract a large number of test-takers might not be clear and need to be reviewed.

Table 4. *Item distracters*

No	% of A's	% of B's	% of C's	% of D's	No	% of A's	% of B's	% of C's	% of D's
1	0	90.6	6.25	3.13	11	3.13	87.5	3.13	0
2	0	90.6	9.38	0	12	0	0	100	0
3	9.38	81.3	3.13	0	13	12.5	15.6	9.38	62.5
4	25	28.1	40.6	6.25	14	18.8	40.6	21.9	12.5
5	0	6.25	68.8	18.8	15	6.25	0	87.5	0
6	56.3	15.6	25	3.13	16	12.5	15.6	15.6	46.9
7	78.1	15.6	3.13	0	17	15.6	25	21.9	25
8	12.5	65.6	9.38	12.5	18	9.38	37.5	18.8	18.8
9	6.25	81.3	9.38	0	19	25	3.13	40.6	21.9
10	0	50	46.9	0	20	43.8	18.8	6.25	15.6

The table above indicates that the question items have several types of distractors. Firstly, there are 9 items which have one or two distractors that were never chosen by test takers, such as the item number 1, 2, 3, 5, 7, 9, 10, 11, and 15. Secondly, there is one item (the item number 6) which has a distracter and was chosen by many test takers (option A, 56.3%). Lastly, all distractors in the item question number 12 were never selected by test takers since the question item is arguably quite easy.

Revision of test items

Based on analysis of the question items, the Cronbach's alpha (α) of the tests is 0.79; it can be concluded that the test items are reliable. However, after analyzing the facility value and discrimination index of each item, we have decided to revise two items, which are most difficult item and the easiest one (the items which do not have a discrimination index value).

The revised items occurred in the item number 6 and 12. Item number 6 has facility value of 0.03, which indicates that it is quite difficult. As a result, we have revised the multiple-choice options only. Meanwhile, for item number 12, we have revised the question element of the test item since the item facility value is too easy (IF = 1). The revisions made are as follows.

Revision 1 (multiple choice)

6. Mengapa Museum Bahari masih ditutup pasca kebakaran?
(Why was the Museum of Bahari still closed after the fire?)
(A) karena hanya akan ada kegiatan bersih-bersih
(because there will only be clean-up activities)
(B) karena di bagian dalam masih terpasang garis polisi
(because the area inside is still applied the police line)
(C) karena akan ada investigasi lanjutan dari pihak kepolisian
(because there will be further investigation from the police)
(D) karena untuk kepentingan penyelidikan dan pengamanan dari warga
(for the purposes of the investigation and security from people)
(The correct answer is D, but the most chosen answer is A, 56.3%)

The item number 6 is considered as very difficult item (IF = 0.03). Moreover, based on the result of distractors item analysis, most test takers have chosen the option A as the answer which is incorrect. After reviewing the item, we can conclude that the item difficulty is due to misidentification of the location in the paragraph where the correct answer lies. Thus, we modified option A. Below is the revision of the answer in item number 6.

(A) Untuk mencegah warga masuk ke area kebakaran
(to prevent people from entering the area of the fire)

Revision 2 (whole item)

12. Kapan rencana Museum Bahari akan dibuka kembali?
(When will the Museum Bahari plan be reopened?)
(A) 16 Januari mendatang
(B) 17 Januari mendatang
(C) 19 Januari mendatang
(D) 20 Januari mendatang

Item number 12 is considered as a very easy item (IF=1), and has no discrimination value, nor any distraction item. After reviewing the item, this is caused by the information of the answer being obvious in the text. Therefore, we made a total revision in the question. Here is the total revision for item number 12.

12. Area mana saja yang akan dibuka pada Jumat 19 Januari mendatang?
(Which areas will be open on Friday 19 January?)
(A) gedung yang terbakar
(the burnt building only)
(B) gedung yang tidak terbakar

(unburnt building only)

(C) *semua area, khususnya gedung yang tidak terbakar*
(all areas, especially unburnt buildings)

(D) *semua area, kecuali gedung yang terbakar*
(all areas, except the burned building)

(The answer is C)

Discussion

In terms of validity of the content, it can be concluded that the content of the test is valid and appropriate because it contains three different types of authentic Indonesian texts, such as exposition, news, and literature texts. In addition, the descriptive analysis indicates that the score of students' distributions are equally at beginner, intermediate, and advanced levels. Although students are in the same class, Indonesian 3, the students' background experience in learning Indonesian is varied, such as less than 1 year, 2 to 5 years, and more than 6 years.

The item facility value analysis shows that the test items developed already have a balanced difficulty: easy, moderate, and difficult. However, there is one item considered very easy, and one very difficult. In addition, we also found a good discrimination index value in the test items that we have designed. However, there is one item that did not appear in SPSS, because it has 0 value of discrimination index. Based on the analysis of item facility and discrimination index values, we have revised to the two question items, item number 6 and 12. The item that contributes to the students' failure can be considered as quite a difficult item for the test takers. It could be argued that question 6 assesses the test takers' ability to make an inference from the text, provided that the answer is not explicitly mentioned in the text. However, in the distractors, there is one option (option A) that has a close relationship to the question. Hence, students may think that the answer is the aforementioned distractor. Meanwhile, question 12 contains an obvious answer which requires test takers to choose the date of an event. This is very easy since the answer option mentioning the event date is clearly matched with that written in the text.

Furthermore, the findings in this study also have implications for the discussion of the format of reading tests. The study refutes the Pyrczak's (1975) argument which states that the use of multiple-choice type questions leads to a lack of dependence on reading the passage on the part of test takers. In fact, in this test material, students need to read the passage to find answers that match with the information and context. Moreover, the study also supports the findings of Baghaei and Ravand (2015) which suggest that multiple choice formats can trigger cognitive processes and learners' comprehension to choose the most appropriate answer. Furthermore, regarding its characteristics, the reading test designed is also in line with the development theory of tests by Hughes (2003) and Brown (2004), which describes standardized stages.

In addition, this study also supports the findings of Boyaci and Guner (2018) which states that the use of authentic materials has an impact on students' reading comprehension and also has positive responses from students. This test provided authentic material by way of three different types of text. Furthermore, with varying levels of test difficulty, this test also supports the argument of Wilson (2007) suggesting that the text of reading should be challenging so that students can feel achievement in answering the test. Regarding BIPA teachers, this test also complements the finding of Kamgar and Jadidi (2016), related to the contributions for foreign language teachers when developing evaluation tests. To evaluate their students, teachers of *Bahasa Indonesia* can prepare questions items that refer to the standards contained in the instruments developed in this study.

Conclusion

The development of tests to measure proficiency in *Bahasa Indonesia* for NNS is necessary. This research focuses only on the aspect of reading comprehension with multiple choice formats because the characteristics of this type of test are commonly used for the large-scale testing. In this study several steps have been undertaken, such as designing tests, piloting, administering tests, analyzing test items, and revision based on the results of analysis. This test can be used by institutions to measure the level of proficiency in reading of NNS of *Bahasa Indonesia*. In addition, this test is also likely to be used by teachers and universities to determine student class placement in a class for university. Development of reading tests with a greater number of questions, for example, 40 items would also appear warranted.

In this study, a set of test items for Indonesian language proficiency was developed through a piloting process. The content validity shows that the questions are valid and reliable. However, no statistical procedures have been undertaken to measure the validity of the items during the piloting process. As a result, we have found that some items need to be revised after we administered the test. Based on calculations of facility value, one question contributed to the test takers' failure. In addition, one question was considered as a very easy question and which all test takers could answer correctly. Thus, to create a more reliable test, it is recommended that during piloting the item not only be evaluated by seeking feedback from experts but also by undertaking statistical measurement.

References

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Baghaei, P., & Ravand, H. (2015). A cognitive processing model of reading comprehension in English as a foreign language using the linear logistic test model. *Learning and Individual Differences*, *43*, 100-105.

Belet Boyaci, S. D., & Güner, M. (2018). The impact of authentic material use on development of the reading comprehension, writing skills and motivation in language course. *International Journal of Instruction*, *11*(2), 351-368.

Bernhardt, E. (1983). Testing foreign language reading comprehension: The immediate recall protocol. *Die Unterrichtspraxis / Teaching German, 16*(1), 27-33.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge: Cambridge University Press.

Brown, J. D. (2004). *Language assessment: Principles and classroom practices*. White Plains: Pearson Education, Inc.

Cranney, A. G. (1972). The construction of two types of cloze reading tests for college students. *Journal of Reading Behavior*, *5*(1), 60-64.

Farhady, H. (2012). *Principles of Language Assessment.* New York: Longman Inc.

Gorjian, B. (2013). The effect of passage content on multiple-choice reading comprehension test. *Procedia-Social and Behavioral Sciences*, *84*, 160-164.

Gorsuch, G., & Taguchi, E. (2008). Repeated reading for developing reading fluency and reading comprehension: The case of EFL learners in Vietnam. *System*, *36*(2), 253-278.

Gorsuch, G., & Taguchi, E. (2010). Developing reading fluency and comprehension using repeated reading: Evidence from longitudinal student reports. *Language Teaching Research*, *14*(1), 27-59.

Heaton, J. B. (1988). *Writing English language tests*. New York: Longman Inc.

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

Jones, R. L. (1977). Testing: A vital connection. The language connection: from the classroom to the world. *ACTFL Foreign Language Education Series, 9*.

Kamgar, N., & Jadidi, E. (2016). Exploring the relationship of Iranian efl learners' critical thinking and self-regulation with their reading comprehension ability. *Procedia-Social and Behavioral Sciences*, *232*, 776-783.

Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading*, *12*(3), 281-300.

Nindyaningrum, F. W. (2018). Pengembangan instrumen asesmen uji kemahiran membaca bagi penutur asing. *Thesis, Master Program, Universitas Malang.*

Pyrczak, F. (1975). Passage-dependence of reading comprehension questions: Examples. *Journal of reading*, *18*(4), 308-311.

Rahmiati, I. I., & Emaliana, I. (2017). Developing reading test using lower to higher order of thinking for esp students. *Language in India*, *17*(11). 124-144.

Saifudin, M. F., Suwandi, S., & Setiawan, B. (2014). Pengembangan model tes kompetensi berbahasa Indonesia. *Thesis, Universitas Muhammadiyah Surakarta.*

Sellers, V. D. (2000). Anxiety and reading comprehension in Spanish as a foreign language. *Foreign Language Annals*, *33*(5), 512-520.

Shohamy, E. G. (1981). The cloze procedure and its applicability for testing Hebrew as a foreign language. *Stanford University / University of California, Berkeley,* 101-114

Taguchi, E., Gorsuch, G., Takayasu-Maass, M., & Snipp, K. (2012). Assisted repeated reading with an advanced-level Japanese EFL reader: A longitudinal diary study. *Reading in a Foreign Language*, *24*(1), 30-55.

Taguchi, E., Takayasu-Maass, M., & Gorsuch, G. J. (2004). Developing reading fluency in EFL: How assisted repeated reading and extensive reading affect fluency development. *Reading in a Foreign Language*, *16*(2), 70-96.

Weir, C. J. (1990). *Communicative language testing*. New York: Prentice Hall.

Wilson, K. (2016). Critical reading, critical thinking: Delicate scaffolding in English for academic purposes (EAP). *Thinking Skills and Creativity 22,* 256-265.

Biographical notes

ANDIKA EKO PRASETIYO is a fulltime student at Melbourne University, Master of Applied Linguistics program. He holds a Bachelor of Education, concentration Indonesian Language and Literature Education at *Universitas Negeri Semarang*, Indonesia. His research interest includes Indonesian education, language testing, and educational technology. Email: andikaekop@gmail.com