DEVELOPING READING ASSESSMENT INSTRUMENT FOR INTERMEDIATE EFL LEARNERS: AZWAR MODEL

Fitriatul Masitoh
English Language Education Department, Faculty of Letters, Universitas Negeri Malang, Indonesia
English Language Education Department, Faculty of Tarbiyah, IAIN Kediri, Indonesia
Email: fitriatulmasitoh@iainkediri.ac.id

Ima Fitriyah (Corresponding author)
English Language Education Department, Faculty of Letters, Universitas Negeri Malang, Indonesia
English Language Education Department, Faculty of Tarbiyah, IAIN Kediri, Indonesia
Email: imafitria@iainkediri.ac.id

Nur Mukminatien
English Language Education Department, Faculty of Letters, Universitas Negeri Malang, Indonesia
Email: *nur.mukminatien.fs@um.ac.id*

APA Citation: Masitoh, F., Fitriyah, I., & Mukminatien, N. (2023). Developing reading assessment instrument for intermediate EFL learners: Azwar model. *English Review: Journal of English Education*, *11*(3), 925-934. https://doi.org/10.25134/erjee.v11i3.7461

Received: 27-06-2023 Accepted: 21-08-2023 Published: 30-10-2023

Abstract: The study was inspired by the fact that teachers of an intermediate English reading course employed a pre-made reading instrument particularly tests obtained from the internet, the TOEFL, or a textbook without consulting the learning objectives. Therefore, they cannot meet the demand of the teaching and learning objectives and the classroom-based assessment that should be tailored to the unique circumstances of the course. The objective of this study is to create intermediate reading comprehension test items for EFL students that adhere to the "good test" criterion generated utilizing the Azwar Model (1996). Second-semester students of English Language department in one of institutions in Kediri, Indonesia were meant to take a reading test using this instrument. The results of the try-out from 75 EFL students from the same level demonstrates that the test met the requirements of being valid, reliable, and practical, had moderate difficulty level, good discrimination levels as the most, and 73 functional distractors of the total. Meanwhile, five-invalid test items are not going to be included to be utilized in assessing EFL students' reading comprehension. Intermediate reading course lecturers could then utilize the 35 items which are considered as valid, reliable, and moderately difficult to measure their students' reading achievement.
Keywords: *item analysis; reading test development; reliability; validity.*

INTRODUCTION

Learning to think critically is one of the many benefits that reading provides. It is a thought process that needs quiet time for investigation (Al Roomy, 2022; Younis et al., 2023) Its activities assist the reader to get a deeper comprehension of and respond to the material being read. Readers do a number of things to understand what is written, understand the context of what they are reading, judge the quality and value of the information, and make decisions about what to read. Putting reading assignments in the right order takes some thought. Reading can be broken down into different levels based on how much thinking is involved. In 1996, Burns, Betty, and Ross created a reading taxonomy with four levels: literal, interpretive, critical, and creative. Literal reading means that the readers are able to get information directly from what is said. In order to do interpretive reading, you have to be able to figure out what information is implied by the interline statement. In critical reading, the readers are able to learn new things by using critical thinking skills. For creative reading, you need to be able to imagine things and be creative in order to come up with ideas.

Reading and comprehension may appear to be distinct concepts, but they are actually two components of a larger whole that need continual progress in education. Reading comprehension is the expected outcome of reading and is described as the ability to combine prior knowledge with reading materials (Joh & Plakans, 2017) The readers' experiences, abilities, motivation, and reading goals influence their level of comprehension (Kuşdemir & Bulut, 2018). According to Grabe and Stoller (2019), it entails recognizing and comprehending the fundamental

concepts of texts and drawing inferences based on both texts and prior knowledge. Thus, reading comprehension is essential for lifelong learning. To help students learn how to comprehend, Davis and Vehabovic (2018) say that tests have become more important in recent years because they are used to evaluate students' progress. Most of the time, teachers use multiple-choice tests to see how well their students are doing. The results of a reliable and valid test can be used by a lot of people quickly and for a low cost, and they can be repeated endlessly (Shohamy, 2020). Because of this, a number of studies have been undertaken on the development of reading comprehension tests (Evenddy et al., 2021; Hanafi, 2016; Ozdemir & Akyol, 2019; Perkasa, 2020).

Clearly, classroom-based reading assessment approaches that are successful, suited for EFL classroom demands, and easily applicable in classroom instruction are required. In assessment process, every learning objective should be assessed to determine the outcome of the learning success. Concomitantly, a test is the most frequent sort of assessment instrument. According to Brown and Lee (2001), a test is an evaluation that can provide authenticity, motivation, and feedback to students. The test can be designed to provide students with a score, enabling them to assess and enhance their abilities. Tests, according to Lee (2017), have lately grown in importance due to their role as formative assessments in supporting students' learning. According to Brown (2004, p. 3), the test is "a method of measuring a person's ability, knowledge or performance in a given domain". The test results in student achievement in the teaching and learning process on a regular basis. In other words, the test is meant to evaluate the skill, ability, and knowledge of the students. Teachers form judgments about the nature of a student's reading based on a sample of reading behavior while assessing reading (Boubris & Haddam, 2020). When preparing to conduct assessments, most educators should be aware of the necessity to use established psychometric criteria for determining the reliability and validity of quantitative measures of language and fundamental reading skills (Fitriyah, et al., 2022)

Unfortunately, several of the reading tests models utilized by teachers are not well-suited for assessing reading in the classroom. Preliminary findings suggest that the majority of reading tests administered by reading teachers are not for classroom-based reading test, such as the TOEFL reading test, reading test from internet, and reading tests taken from text book. It also lacks systematic, thorough, and appropriate reading assessments in the classroom. The purpose of this study is to give students with a more authentic reading test development model that will allow them to truly measure their reading abilities in the EFL classroom environment. Numerous factors must be considered when developing tests. The most critical characteristics are reliability, validity, and item analysis. Considering the importance of these, this study also purposes at developing a valid and reliable reading comprehension test based on *Merdeka Belajar* Curriculum in higher education for undergraduate students of Intermediate Reading course. It also due to the fact that the department of English Language Education in one of Islamic institutions in Kediri has not prepared clear course description for the intermediate reading level in Intensive Reading Course. This study's findings will ideally fulfill the institution's need and others for a ready-made reading test for assessing students at the intermediate level, or the so-called Intensive Reading course in the *Merdeka Belajar* Curriculum.

Previous studies on the construction of a reading test have shown some interesting results. Hanafi (2016) developed a reading test for the first semester students of university level and was intended to be used as instrument for research in reading. In addition, Ozdemir and Akyol (2019) did another study in which they designed a valid and accurate reading comprehension test for fourth grade children. Furthermore, Azmi (2020) conducted research and designed a web-based reading comprehension test for second semester English Language Education students. Finally, Ningrum and Sudarwati (2022) created an Indonesian EFL critical reading test. Despite the Critical Reading test's validity and reliability, students' tryout scores are poor.

Traditional reading comprehension tests have been widely used for decades, typically focusing on assessing a reader's ability to understand and recall textual information. While these assessments serve as valuable tools for evaluating a reader's comprehension skills, they often fall short in capturing the full complexity of reading comprehension. One significant gap in current reading assessment practices is the limited attention given to higher-order cognitive processes such as critical thinking, inference-making, and text analysis. The conventional assessments primarily measure what readers know rather than how well they can apply their

p-ISSN 2301-7554, e-ISSN 2541-3643
https://journal.uniku.ac.id/index.php/ERJEE

knowledge to think critically and draw insightful conclusions from a text. Thus far, the study concerning the development of Intermediate Reading test was rarely done. Thus, in addition to solving the problem in Intermediate Reading class due to the condition of the absence of clear course description and inappropriate reading test as what have been previously described in the preliminary study, this study tried to fill a gap that other studies had left by determining the ideal kind of intermediate reading test for English majors in colleges. This study focused on the development of an intermediate reading test as an assessment tool for EFL students in Intensive Reading course.

METHOD

The primary objective of this project is to develop intermediate reading assessments for second-semester English Language Education Department students in one of Islamic institutions in Kediri. This is a study of research and development (R&D). Research and development is a product development model in which results are used to produce new solutions (Gall et al., 2003). Seventy-five students were selected throughout the try-out phase as student participants. In the peer debriefing phase of the validation procedure, one instructor of Intermediate Reading courses and a developer of language tests participated.

In accordance with the study's primary purpose, the model of development used in this research is Standard Step of Developing a test, adapted from Azwar (1996) deemed to be complete procedures as there are nine steps proposed in this model as showed in the Figure 1.

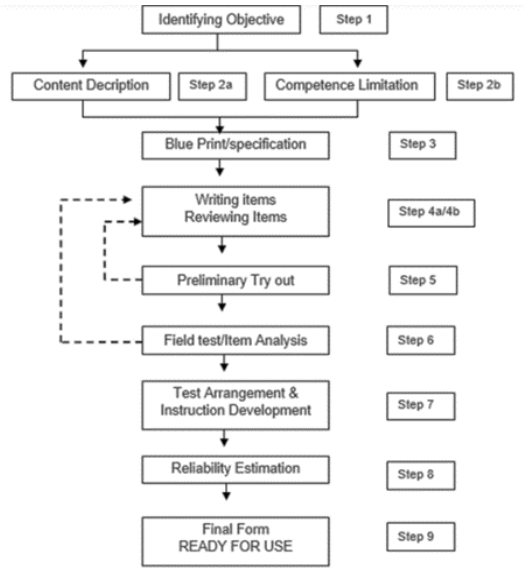

Figure 1. *Step of developing a test* (Azwar, 1996)

Initially, a pilot study was conducted to determine the significance of developing an Intermediate Reading test. The process continued with the test development, which involved the following steps: developing the test based on the course description, determining the objective based on the course description, producing the test item, checking and rechecking the test item, and validating the test (peer debriefing). The next step was to do a try-out test, look at the results, and evaluate the difficulty level of the multiple-choice questions and how reliable they were. The method is depicted in the diagram below.

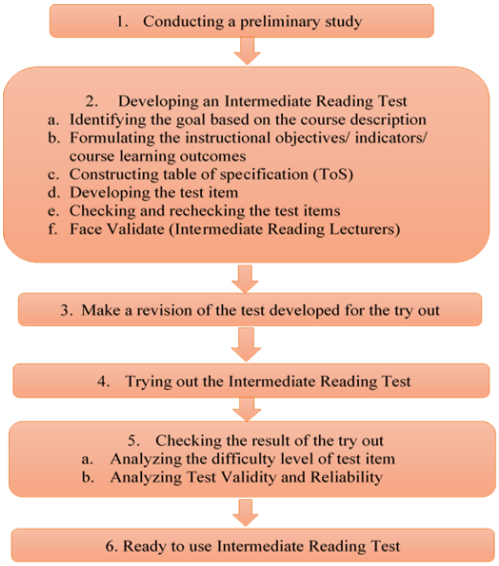

Figure 2. *The stages of intermediate reading test development*

Then, the test's indicators could be made based on its goals and objectives. The developers developed a reading test as the final product. The test consists of 40 multiple-choice questions with four possible answers. In the try out phase, 75 English department students, approved by ethics committee of the institution, almost completed intermediate reading course were invited as participants. After trying out, the developers used item validity to evaluate the test. The item validity test analyses each item on the reading test using Point-Biserial Correlation. For Brown (2004), validity refers to how well a particular method of data gathering yields the desired results. Data from a reading test were analyzed with an emphasis toward the test's validity and reliability. This analysis looks at how well students can take a reading test and how good the test is. To finish, the test's reliability and internal consistency were assessed by using the Kuder Richardson-20 formula (K-20). Statistically, a number between 0 and 1.00 will indicate the degree to which the

scores related. A score near 1.00 shows high reliability, while a score closer to 0 shows low reliability. The final result will contain the items which are valid and have high reliability.

RESULTS AND DISCUSSION

The findings built in a systematic manner using the aforementioned approaches concerned with a chronological description of test development techniques. Here is the outcome of this R&D study.

A preliminary study was done firstly. The urgency of establishing an ideal form of Intermediate Reading test for second semester students was based on the results of a preliminary study conducted before the semester began in response to the evaluation of the Intermediate Reading course that one of the authors taught previously. Based on the observation and interview with the lecturers, most the reading assessment instruments were taken from TOEFL book. They admitted that the objectives were different from the course objectives. Moreover, the curriculum also changes from the KKNI to *Merdeka Belajar* Curriculum recently. The department does not provide clear course description, since each lecturer of Intermediate Reading Course makes the course description by themselves. Therefore, because of this condition, seems that the lecturers themselves are not in one idea in teaching the course. There should be a clear guidance for them that could lead to the best outcomes.

Developing Intermediate Reading Test is the next step of the development. This project was carried out to develop reading test items for undergraduate students of Intermediate reading course as a final test. Test item development began with a course description; instructional objectives; indicators; and course learning outcomes were formulated. Then, test items were developed, and the test items were checked and rechecked. Finally, the test items were put to the test and the results of the test were examined.

The test was developed by referring to course description of the Intermediate Reading Course profile in the English Education Study Program, Universitas Negeri Malang. The course profile was chosen because it provides complete guidance for the intermediate reading level. Thereby, the goal of the test is that the students are able to understand short academic articles and story by applying reading strategies in identifying keywords, making inferences, analyzing dictions, interpreting culture-bound dictions, analyzing organization and development of ideas, and identifying text types effectively.

Indicators of the test are: (1) identifying keywords: identify topic of the text and the meaning of difficult word from context; (2) making inferences: answer inference questions; recognize stated or implied meaning; (3) analyzing dictions: recognize clues (synonym / antonym); (4) interpreting culture-bound dictions: recognize specific term; (5) identifying text patterns: identify listing, time order, cause-effect, and comparison-contras; and (6) analyzing organization and development of ideas: being able to distinguish between the main idea, topic sentence, supporting details, and conclusion.

In accordance with the blueprint, the reading exam as the product was developed. Reading test passages were culled from publications aimed at intermediate readers and the websites. This is a 40-item test with four response possibilities. Correct answers receive a score of 1, while erroneous answers receive a score of 0. It is worth one point for each correct answer, and 0 points for each bad answer.

The next step is to make sure that the product is good. Peer debriefing was then used to check the quality of the product made in this step. Peer debriefing, also called "analytic triangulation," is the process by which a researcher gets in touch with a peer who is not involved in the research to help him or her figure out what the researcher thinks about all or part of the research process. A reading lecturer of an Islamic institute of Kediri (debriefed peer 1) and the head of the East Java English MGMP and a national English test writer (debriefed peer 2) were among the peers involved. Although the developers have experiences in creating a number of assessment instruments, the input of those who will utilize instrument is crucial. The debriefed Peer 1 concerned on the item 1 and 7, which are nearly identical. Then we make a slight modification in order to reduce the similarity. The next is item number 3, the distractor, is extremely perplexing because the responses to the distractors are nearly identical. During the debriefing, peer 2 raised the choice of response, which should be altered with regard to the level of difficulty it presents, and the influence of distractors. He remarked that the questions appeared to be relatively simple for the student level; they were nonetheless understandable for lower-level students. For example, a text about a comparison between university and school culture seem so easy for their level, he suggested omitting the text. We agreed not to put the text and to find

p-ISSN 2301-7554, e-ISSN 2541-3643
https://journal.uniku.ac.id/index.php/ERJEE

more pertinent material.

After getting feedback from peers during a "peer debriefing" session, the product was changed. Based on what was talked about at the peer debriefing session, the researchers made some changes. The researcher changed a few test items from easy to moderate so that they would be the best type of questions. The study also found the best way for students to improve their ability to think critically while doing something else. Students will be able to use their critical thinking skills to their fullest by using distractions well. Since this is done online, with less supervision, the time was made up in a way that makes it less likely that students will lie. Testing the product was the next step. In this step, the researchers did a test to see how valid and reliable the test was.

Next, the test's reliability was investigated using the Kuder Richardson-20 formula, validity with the Point-Biserial Correlation formula, and item analysis. According to Danuwijaya (2018), item analysis is both a conscious and unconscious process that regularly evaluates the quality of each item. It is helpful to find difficult and easy options, to examine how the function discriminates between low and high scores, to alternate the function, and to compile a solid question bank. The outcome of this phase serves as the foundation for the subsequent phase.

Reliability

Reliability, the capacity of assessment instruments to generate steady and consistent outcomes (Hughes, 2003), measurement of the intermediate reading test was done through an analysis that is by Kuder Richardson 20 as the internal consistency. The result of internal analysis consistency is presented in Table 2. The overall test instrument analyzed with the data reliability KR-20 in point 0.851. These results indicate the condition of the reliability of KR-20 included in the category of reliable as a form of internal consistency.

Table 1. *Reliability test of test items*

KR − 20	K	Note
.851	40	Reliable

Validity

To test the validity, indicating that the reading test should be able to measure what it should measure emerging as the process of obtaining facts to argue for the legitimacy of test interpretations and decisions (Giraldo 2020; Heaton, 1991), of the question items, the researchers used Point Biserial correlation statistics (r_Pbi) with criteria if the value of r_Pbi>r_table (0.227) then the test item is declared valid. The following table summarizes the findings of the test of item validity.

Table 2**.** *Item validity testing*

Category	Number of test item	P(%)
Valid	1,2,5,6,7,8,9,10,11,13,16,17, 18,19,20,21,22,23,24,25,26, 27,28,29,30,31,32,33,34,35, 36,37,38,39,40	87,5 %
Invalid	3,4,12,14,15	12,5 %

Based on the statistic tabulation, five items are considered as invalid test. These items were then excluded to be used in the real test.

Item difficulty level

The proportion of students that properly respond to a specific item is related to item difficulty. Diverse levels of difficulty should be incorporated into the passages to prevent students from experiencing anxiety, tedium, and exhaustion (Masduqi & Fatimah, 2022). The level of difficulty can be determined by evaluating the replies of the students. It indicates that the difficulty of the questions was determined by students' responses rather than teachers' views (Wijayanti, 2020). This research used Heaton's (1990) formula to measure the Facility value (FV) or difficulty level which is gained by dividing the number of students from the upper group and the lower group students who answer a certain item correctly by the total number of the students who join the test. To categorize the FV, the classification from (Djiwandono,1996) used to find the difficulty level of the reading test. The classification is as follows:

Table 3. *Classification of reading test difficulty*

Classification	Interpretation
0.000 – 0.250	Difficult
0.251 – 0.750	Moderate
0.751 – 1.000	Easy

The computation results for the difficulty index are shown in the table below.

Table 4. *Summary of the item difficulty*

Difficulty level	Test item F	%	Number of test item
Difficult (P < 0,250)	3	7,5	3, 4, and 14
Moderate (0,251< P < 0,75)	15	37,5	5, 10, 11, 12, 13, 15, 17, 18, 19, 20, 25, 28, 29, 34, and 38
Easy (P >	22	55	1, 2, 6, 7, 8, 9, 16,

0,751)	21, 22, 23, 24, 26, 27, 30, 31, 32, 33, 35, 36, 37, 39, and 40

The majority of the items in table 5 have an easy level of difficulty. There are 22 questions, or 55 percent of 40 questions are in easy level. Meanwhile, there are 15 questions, or 37,5 percent of 40 questions, with a moderate level and the remaining 3 items fall within the difficult category notably question number 3, 4, and 14. The following examples (Table 5) illustrate each example of a test item for each difficulty level. In addition to this tabulation, the students' scores are various, range from 20 (The lowest score) to 97.5 (the highest score), with the means score is 70.9.

Table 5. *An example of a test item from each difficulty level*

Level of difficulty	Test item
Easy	(22) Why did Andrea suddenly stop the car?
Moderate	(11) What is the best title for the text?
Difficult	(4) A broken human-made dam is compared to what?

Item discrimination level

The reading test items' discrimination level was also examined. This can tell which students have learned the subject and which have not. The discrimination index analysis by Arikunto (2008) is used to evaluate item discrimination power, as shown in Table 6.

Table 6. *Item discrimination level*

Discrimination level	Category	F	P	Item number
0,70 – 1,00	Excellent	3	8%	22,25, 28
0,40 – 0,70	Good	17	43 %	5,11,13,17,18,19,20,23,26,27,29,34,35,36,37,38,40
0,20 – 0,40	Satisfactory	11	28 %	1,2,3,6,7,10,24, 30,31,33,39
0,00 – 0,20	Poor	7	18 %	4,8,9,15,16,21, 32
Negative	Rejected	2	5%	12,14
Total		40	100 %	

It can be seen that items with low discriminating power account for only 18% of the total. The table, on the other hand, reveals that products with high discrimination levels are the most common with 43% of the total. Surprisingly, 28% of the discrimination level of the items is satisfactory, while 8% is great. This finding suggests that those things should be somewhat modified. Items that are satisfactory, good, or excellent can be reused and saved immediately in the questions bank.

Distractor analysis

Distractor analysis is utilized to determine the effectiveness of inappropriate options in distracting the lower groups (Manfaat et al., 2021). In this study, each item's detractors have been studied. If lower-level students are more likely to select incorrect responses than higher-level students, an item is a good diversion. In a majority of assessments of learning outcomes, a distractor is seen to have served its purpose well if it has been chosen by at least 5 percent of test-takers (Sudijono, 2009). Nonetheless, if 1% to 4% of test-takers have selected them, they must be updated (Daryanto, 2005). The result of the analysis of distractor effectiveness is displayed in the table below.

Table 6. *Distractor analysis*

Distractor efficiency	Category	F	P (%)	Item number
0	Rejected	7	6%	1,8,16,19,31, 32,35
1-4%	Revised	40	33%	1,2,,5,6,7,8,10,11,12,16,17,18,19,20,21,22,23,34,26,28,30,31,32,33,34,37,38,39,40
5%	Accepted	73	61%	1-40
Total		120	100%	

It can be noticed from the table that 73 distractions are functionally sound. In contrast, just seven distractors are rejected, whereas forty require modification. Possibly, the only flaw resides in the formulation of the sentences; therefore, it merely needs to be recast with the required modifications. Writing questions is tough, therefore if they can be fixed, they should be fixed rather than discarded.

This study aims to develop multiple-choice intermediate reading questions that are in accordance with a clear course description and provide reading questions that are valid, reliable, have a moderate level of difficulty, good

discrimination level, and acceptable functional distractors. The findings show that the Intermediate Reading test with 40 items have been developed based on the proper procedure, Azwar Model (Azwar, 1996) and have called as reliable, but it is quite unfortunate with the large number of try out participants, five questions were invalid, therefore these five questions will not be used for the actual test. Comparing the present study to earlier research (Azmi, 2020; Hanafi, 2016; Karim & Haq, 2014; Ningrum & Sudarwati; 2022; Ozdemir & Akyol, 2019;), there are some similarities and differences. The similarities lay in the fact that both the prior research and the current study are able to provide a reliable test for usage in the field requiring the reading test test. The difference is that in the present study, although the student performance is in moderate level, some of the items were invalid. Despite the fact that five items were deemed invalid, all of the items were reliable, and the results of the test session were largely as planned, as some students received high scores.

Finding related to validity in this study are slightly different from Sudarwati et al.'s study (2021). Their study found that the multiple-choice questions that had been tested on 50 students showed that all of the items were valid and had met the proportion of questions in the moderate category but the students' scores were at a low level. Meanwhile, this recent study shows different results. The questions that were tested on larger number of (75) students resulted in quite varied scores, with a fairly good average score of 70.9 even though there were some questions that were not valid. There are several potential outcomes for the occurrence. The first is related to the test administration (Brown, 2004). This test was done online so that there is no direct supervision. In addition, a longer duration of the time in doing test may also have an effect on the difficulty level of the questions, which this study shows that the majority of the test questions are easy. The second issue is student preparation. This contradicts the findings of Ellis and Ryan (2003) that discovered that students' low-test scores were related to inadequate preparation. According to them, students' lack of preparation before taking an exam may result in a lower score. Even though the students in this study were not aware that they would be taking a test, the mean score of 70.9 indicates that they performed pretty well.

The difficulty level also indicates that, on general, the questions are moderately challenging for students in the second semester as they offered multiple levels to avoid student anxiety, boredom, and exhaustion (Masduqi & Fatimah, 2022). Furthermore, the average score achieved by the students is a respectable 70.9, which contradicts the findings of Evenddy et al. (2021). According to their research, 85 percent of students in higher education consider the vocabulary query category to be hard. This is possible since this test was administered to students who had mastered the majority of the course's subject, so they could do the test well. The absence of direct supervision by the teacher is another possible cause of this outcome. The reading test was administered asynchronously using Google Form; despite the limited time, direct supervision is necessary in an evaluation to prevent unfair student behavior (Fitriyah & Jannah, 2021).

Concerning the item discrimination index and the efficacy of distractor analysis as part of the item analysis of the reading test for students in the English department, the present study is more comprehensive than the majority of relevant studies that didn't do any testing (Azmi, 2020; Ningrum & Sudarwati, 2022; Sudarwati et al., 2021). In addition, this study is a compliment to those that examined three indicators of item analysis, namely item difficulty, item discrimination, and distractor analysis, in terms of having more reading test items (Hanafi, 2016) and the number of English department test takers (Danuwijaya, 2018).

CONCLUSION

This study aimed to develop a reading test for intermediate-level EFL students in Indonesia. Researchers came up with a set of test development procedures that were used to make the product. These steps are completed in the following order: determining the goal based on the course description, developing instructional objectives/indicators/course learning outcomes, developing the test item, checking and rechecking the test item, administering a practice test, reviewing the results, and assessing the difficulty level and reliability of the test items. The analysis shows that the biserial correlation value for all items is higher than 0.227, which means that almost all items meet the validity criteria (5 are invalid). The analysis also showed that the KR-20 value is0.851, which is higher than 0.70. This means that all inquiries about items meet the criteria for reliability. Because of this, one could say that the Intermediate Reading test is good for the intensive reading course. In terms of how

difficult the test was, most of the questions were found to be moderately difficult. There are three difficult questions, fifteen moderately difficult questions, and twenty-two easy questions. Concerning discrimination levels, the findings show that 8% of item questions are excellent, while those with a high discrimination level account for 43% of the total, and 28% of the discrimination level of the items is acceptable. Furthermore, 73 distractions were found to be functionally sound after being tested. In comparison, just seven detractors are rejected, while forty demand modification. Based on this, the test is suitable for usage. In addition, this study shows that the validity of the product has met 87,5% of 40 reading test items; nonetheless, the reliability standards and the exam have been shown to be effective in assisting students in learning the intermediate course. Even if the exam is reliable, the results of the test are unsatisfactory in terms of the students' performance. This may be due to the absence of teacher supervision and the extended time allotment, which prevents students from completing the test in an optimal and fair manner. As a result, future researchers should conduct field tests over the course of a semester, for instance, by designing an intermediate reading test with clear time allocation and supervision, despite the fact that it is administered online. This can be utilized to address the study's limitation.

ACKNOWLEDGEMENT

The authors would like to express their gratitude to Ibu Yogi Rohana for her time and willingness to participate in the data gathering process, as well as to the Indonesia Endowment Fund for Education (LPDP).

REFERENCES

Al Roomy, M. A. (2022). Investigating the effects of critical reading skills on students' reading comprehension. *Arab World English Journal, 13*(1), 366-381.

Anastasiou, D. & Griva, E. (2009). Awareness of reading strategy use and reading comprehension among poor and good readers. *Elementary Education Online*, 8, 283-297

Azmi, U. (2020). Developing web-based reading tests for the students of English language education. *Journal of Applied Linguistics, Translation, and Literature, 1*(2), 92-104.

Azwar, S. (1996). *Tes prestasi: Fungsi dan pengembangan pengukuran prestasi belajar* (2nd edition). Pustaka Pelajar.

Borg, W.R. & Gall, J.P. (2003). *Educational research: An introduction* (7th Edition). New Jersey Pearson Education.

Boubris, A. A., & Haddam, F. (2020). Reading assessment: A case study of teachers' beliefs and classroom evaluative practices. *Arab World English Journal, 11*(4), 236-253.

Brown, H. D., & Lee, H. (2001). *Teaching by principles: An interactive approach to language pedagogy*. Prentice Hall Regents Englewood Cliffs, NJ.

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

Burns P.C., Betty, D. & Ross, E. P. (1996). *Teaching reading in today's elementary schools.* Houghton Mifflin Company.

Daryanto. (2005). *Evaluasi pendidikan.* Rineka Cipta

Danuwijaya, A. A. (2018). Item analysis of reading comprehension test for post-graduate students. *English Review: Journal of English Education, 7*(1), 29. https://doi.org/10.25134/erjee.v7i1.1493

Davis, D. S., & Vehabovic, N. (2018). The dangers of test preparation: What students learn (and don't learn) about reading comprehension from test-centric literacy instruction. *The Reading Teacher, 71*(5), 579-588.

Day, R. R., & Park, J. S. (2005). Developing reading comprehension questions. *Reading in A Foreign Language, 17*(1), 60-73.

Djiwandono, M.S. (1996). *Test bahasa dalam pengajaran*. Institut Teknologi Bandung.

Ellis, A. P., & Ryan, A. M. (2003). Race and cognitive-ability test performance: The mediating effects of test preparation, test-taking strategy use and self- efficacy. *Journal of Applied Social Psychology*, 33(12), 2607-2629

Evenddy, S. S., Nurlely, L., & Marfu'ah, M. (2021). Reading comprehension test and its challenges in students' perspective. *Loquen: English Studies Journal*, *14*(1), 40-47. https://doi.org/10.32678/loquen.v14i1.2734

Fitriyah, I., & Jannah, M. (2021). Online assessment effect in EFL classroom: An investigation on students and teachers' perceptions. *Indonesian Journal of English Language Teaching and Applied Linguistics, 5*(2), 265-284.

Fitriyah, I., Masitoh, F., & Widiati, U. (2022). Classroom-based language assessment literacy and professional development need between novice and experienced EFL teachers. *Indonesian Journal of Applied Linguistics, 12*(1), 124-134.

Gellert, A. S., & Elbro, C. (2013). Cloze tests may be quick, but are they dirty? Development and preliminary validation of a cloze test of reading comprehension. *Journal of Psychoeducational Assessment*, 31(1), 16-28.

Giraldo, F. (2020). Validity and classroom language testing: A practical approach. *Colombian Applied Linguistics Journal*, 22(2), 194-206.

Grabe, W., & Stoller, F. L. (2019). *Teaching and researching reading*. Routledge.

Grabe, W. 2009. *Reading in a second language: moving from theory to practice.* Cambridge University Press.

Hanafi, H. (2016). Developing reading comprehension test for the first semester students of English department. *ELLITE: Journal of English Language, Literature, and Teaching*, *1*(1). https://doi.org/10.32528/ellite.v1i1.164

Heaton, J.B. (1990). *Writing English language tests.* Longman

Huang, T. W., & Wu, P. C. (2013). Classroom-based cognitive diagnostic model for a teacher-made fraction- decimal test. *Educational Technology & Society*, 16(3), 347–361

Hughes, A. (2003). *Testing for language teachers* (2nd Edition). Cambridge University Press

Joh, J., & Plakans, L. (2017). Working memory in L2 reading comprehension: The influence of prior knowledge. *System*, *70*, 107-120.

Karim, S., & Haq, N. (2014). The process of developing an academic reading test and evaluating its authenticity. *Journal of Language Teaching and Research*, *5*(2), 471–475.

Kuşdemir, Y., & Bulut, P. (2018). The relationship between elementary school students' reading comprehension and reading motivation. *Journal of Education and Training Studies*, *6*(12).

Lee, I. (2017). *Classroom writing assessment and feedback in L2 school contexts*. Springer. https://utpjournals.press/doi/10.3138/cmlr.599

Manfaat, B., Nurazizah, A., & Misri, M. A. (2021). Analysis of mathematics test items quality for high school. *Jurnal Penelitian dan Evaluasi Pendidikan*, *25*(1), 108-117. http://dx.doi.org/10.21831/pep.v25i1.39174

Masduqi, H., & Fatimah. (2022). Assessment in Indonesian higher education: Developing a reading comprehension test for English students. *International Research-Based Education Journal*, *4*(1), 1–13.

Ningrum, A. S. B. ., & Sudarwati, E. (2022). Is the test sensible? Developing a critical reading test for Indonesian EFL learners. *JEELS (Journal of English Education and Linguistics Studies)*, *9*(1), 209–227. https://doi.org/10.30762/jeels.v9i1.4005

Nuttal, C. (2005). *Teaching reading skill in foreign language.* Macmillan.

Ozdemir, E. C., & Akyol, H. (2019). The development of a reading comprehension test. *Universal Journal of Educational Research*, *7*(2), 563–570. https://doi.org/10.13189/ujer.2019.070229

Perkasa, A. N. (2021). Developing and validating EFL reading test. *Education of English as Foreign Language Journal (EDUCAFL)*, *4*(1), 9-18. Doi: 10.21776/ub.educafl.2021.004.01.02

Royer, J. M. (2001). Developing reading and listening comprehension tests based on the Sentence Verification Technique (SVT). *Journal of Adolescent & Adult Literacy*, 45(1), 30-41.

Ranjbaran, F., & Alavi, S. M. (2017). Developing a reading comprehension test for cognitive diagnostic assessment: A RUM analysis. *Studies in Educational Evaluation*, 55, 167-179.

Shohamy, E. (2020). *The power of tests: A critical perspective on the uses of language tests.* Routledge.

Sudarwati, E., Fatimah, F., Astuti, Y., & Ubaidillah, M. F. (2021). Developing online learning assessment instrument for English sentence structure course during covid-19 pandemic. *Langkawi: Journal of The Association for Arabic and English*, *7*(2), 170-181. https://doi.org/10.31332/lkw.v7i2.3122

Sudijono, A. (2009). *Pengantar evaluasi pendidikan*. Rajawali Pers.

Wijayanti, P. S. (2020). Item quality analysis for measuring mathematical problem-solving skills. *AKSIOMA: Jurnal Program Studi Pendidikan Matematika*, 9(4), 1223–1234. https://doi.org/https://doi.org/10.24127/ajpm.v9i4.3 036

Yao, M., & Souza, T. (1992). Developing and Validating an Intermediate English reading test for English as a second language learners. *Report research technical 143*.

Younis, S., Naeem, S., Ali, Z., Yaqoob, N., & Ullah, N. (2023). A study of the relationship between critical reading and critical thinking abilities of undergraduate learners. *Journal of Positive School Psychology*, 1639-1647.