Assessment and Evaluation on Text Readability in Reading Test Instrument Development for BIPA-1 to BIPA-3

Laili Etika Rahmawati[1*], Yunus Sulistyono[2]

[1]Indonesian Language and Literature Education, Faculty of Teacher Training and Education, Universitas Muhammadiyah Surakarta, Surakarta 57102, Indonesia
Laili.Rahmawati@ums.ac.id
[2]Indonesian Language and Literature Education, Faculty of Teacher Training and Education, Universitas Muhammadiyah Surakarta, Surakarta 57102, Indonesia
Yunus.Sulistyono@ums.ac.id
[*]Corresponding Author

https://doi.org/10.24191/ajue.v17i3.14522

Received: 23 May 2021
Accepted: 20 July 2021
Date Published Online: 31 July 2021
Published: 31 July 2021

Abstract: Nowadays, text readability is of great importance. Simple but very often ignored, readability statistics can provide information about the level of difficulty of the readability of particular documents and increase an evaluator's credibility. Hence, this research aims to examine the readability index of the test instrument for BIPA (Bahasa Indonesia Untuk Penutur Asing), evaluate the student's reading ability, and analyze the relevance readability index and cloze test result. The study was carried out through an experiment involving 21 international students in several Muhammadiyah Universities. The students were provided with three sets of reading instruments from BIPA-1 to BIPA-3. Data analysis was carried out with correlation analysis. The result showed that the reading instrument difficulty was at a fairly easy to moderate level. Gunning Fog score and Automated Readability Index were the most relevant tool to the student's test achievement. This research implies the importance of test instrument evaluation. Assessment of the readability of the text is vital in the process of developing an appropriate test instrument.

Keywords: BIPA, cloze, difficulty, instrument, readability.

1. Introduction

A written text is a medium to communicate the writer to the readers (Kasule 2011). Every kind of reading material contains certain information which the reader may need. However, not all of the reading texts can be easily understood. The readability of a text defines if the information it contains is understandable or not. For some cases, the reading materials may be at a higher readability level than the average community's reading ability level (Cheng & Dunn 2015; Sabjan et al., 2021). Thus, the reader may encounter difficulty in understanding the content and extract the information presented.

A second language learner may encounter difficulty understanding a reading text (Xia, Kochmar, & Briscoe 2016). A similar condition may be encountered by the foreign students in Indonesia, including those who join the BIPA program. BIPA (*Bahasa Indonesia bagi Penutur Asing* – Indonesian Language for Foreign Speaker) is a program arranged to improve foreigner's competence in Indonesian (Azizah, Hs, & Lestari 2013). However, currently, there is no distinct knowledge on the readability of test instruments in BIPA.

Measurement of text readability is important to evaluate a document's easiness to read or understand (Kouame 2010). A manuscript with poor readability may lead to the wrong conclusion to

the readers/reviewers. Thus, a text's readability defines whether the article is acceptable or not to the user (Onwuegbuzie, Mallette, Hwang, & Slate 2013: Yilmaz, 2018).

Readability has the hierarchy of difficulty presented as the reading grade. The problem of reading the instrument should be adjusted to the target audience's level (reader) of whom the instrument is developed. The level range is varied among measurement formulas. However, a common understanding is that a higher grade level is related to reading text with a higher difficulty index (Tabatabaei & Bagheri 2013; Binsaleh & Binsaleh, 2021).

Evaluation of the reading text's readability must provide appropriate text quality and comprehensive development stages of reading competence (Xia et al., 2016). On the other side, the readability level also defines the writer's competence in providing target-specific reading difficulty (Zamanian & Heydari 2012). For example, if the reading text's target audience is children, it is supposed to have a high readability level. Otherwise, if the target audience is adults, it is supposed to have a lower readability level. The reading ability of adults is varied among countries. Adult American is considered to have the average reading ability of seventh and eighth-grade level (Kouame 2010), while Australian is considered to be at eighth grade (Cheng & Dunn 2015; Rahmat et al., 2021).

Readability measurement is important to assess the difficulty of a text to understand. There is two kinds of readability measurement methods, such as text-based and reader-based measurement. Currently, there are many tools to measure the readability of a reading text presented as indices, scores, levels, or grades (Alkhurayyif & Weir 2017). Each measurement tool has a unique formula (algorithm) in assessing the difficulty of a reading text. However, not all of the indices are relevant to be used. The relevance is related to various aspects, such as the context of the manuscript (Loughran & McDonald, 2014) or whether the reader is a first language user or second language learner (Zamanian & Heydari 2012; Sabjan et al., 2021).

1.1 Problem of Research

The BIPA program participants are international students who had a particular interest in Indonesian as their second language. To evaluate their learning achievement, an assessment should be carried out. Unfortunately, currently, there is no standardized test instrument for the BIPA program, including reading competence. Reading ability is an aspect that influences someone's awareness (Sultan, Rofiuddin, Nurhadi, & Priyatni, 2017). Thus, an appropriate test instrument is required to ensure the achievement of the learning process. At this point, the higher learning stage should achieve more advanced teaching materials. Thus, there should be improvements to the test quality, especially regarding its difficulty level (Twum et al, 2021).

1.2 Research Focus

The research was focused on the development of a reading test instrument for BIPA. Thus, the result of the assessment should mainly be feedback to the instrument development. This research aimed to examine the readability index of test instruments for BIPA learners at various grades, evaluate the student's ability to understand the text's information and analyze the relevance of text-based index and the cloze test result.

2. Methodology of Research

2.1 General Background of Research

BIPA students are those who learn Indonesian for various purposes. Readability is an aspect of text-based language competence which allows an indirect transfer of information between the writer and the reader. Assessment needs to be carried out to evaluate the whole learning process's outcome (Dunham, Yapa, & Yu 2015). The need for a standardized test instrument for BIPA includes the readability of the reading text instrument. However, the test instrument should also meet the appropriate quality. Particularly for the reading text instrument.

2.2 Sample of Research

The research was carried out at three Muhammadiyah Universities located in several areas in Java, Indonesia. Data collection was carried out between August 2017 and August 2018. The research was focused on the reading text instrument for BIPA. Three levels of BIPA grade were examined during this research. The evaluation involved BIPA learners who joined the Dharmasiswa scholarship program. In total, 21 students met the criteria as the testee (respondent).

2.3 Instrument and Procedures

The data collection method is utilized to fulfill the aim of the study. The materials used in the data collection were the reading text developed for the BIPA test instrument. The format of the reading texts was different from one another. The reading test instrument for BIPA-1 was an advertisement text (source: https://lokermedis.com/tag/perawat/), BIPA-2 was a description of traditional food (source: https://id.wikipedia.org/wiki/Bakwan), and BIPA-3 was a letter. The detailed test instrument form used in this research is shown in **Error! Reference source not found.** to **Error! Reference source not found.** (look at Appendix).

This research is experimental research aimed to evaluate the readability of BIPA test instruments. The instruments were examined from two perspectives, including the difference among levels and the relevance of certain readability indexes. The treatment carried out was the BIPA reading text instruments from level one to level three. Readability indices were calculated to identify the difficulty of reading text instruments. The readability indices include the Flesch-Kincaid Grade Level (FK), Gunning Fog Score (GF), Coleman-Liau Index (CL), SMOG Index (SMOG), and the Automated Readability Index (ARI). The score achievement from the cloze test acted as the dependent variable.

Data collection was carried out through a class experiment. The respondents were provided with three sets of test instruments from BIPA-1, BIPA-2, and BIPA-3. Each instrument consisted of a reading text, and a set of cloze tests consisted of five related questions. The evaluation was carried out for personal achievement and item achievement scores (Cheng, 2020).

2.4 Data Analysis

Data analysis was conducted using ANOVA and correlation statistical analysis techniques. ANOVA was carried out to analyze the difference of the score achievement among the BIPA levels, while correlation analysis was carried out to evaluate the relevance of the readability index to the score achievements. Statistical analysis was carried out with SPSS software, with a confidence level of 90%.

3. Results and Discussion

Analysis of the readability index of test instruments showed the difference of index values. Among the readability index carried out in this research, the GF has the highest index range, while the FK had the lowest value. The detailed calculation result of the readability for the respective index formulation is presented in Table 1.

Table 1. Readability Index of Test Instrument

No.	Instrument	Flesch-Kincaid Grade Level	Coleman-Liau Index	Gunning Fog Score	SMOG Index	Automated Readability Index
1	BIPA-1	15,8	18.8	14.9	9.2	9
2	BIPA-2	18,4	18	21.1	14.5	14.4
3	BIPA-3	16,5	14.8	17.4	12.8	10.6

The indices shown in Table 1 indicates that the test instruments had various difficulty level. Generally, the analysis showed that the difficulty level of reading text instruments is not linear toward

the BIPA grade. The readability index for BIPA-2 is higher compared to BIPA-3 for FK, GF, SMOG, and ARI. However, it changes linearly for the CL.

Each readability index has its own interpretation regarding its level of difficulty. The FK within the range of 15.8 and 18.4 shows the grade of college graduates at all BIPA levels. The readability obtained from the CL was at such a high grade. However, there is no grade category leveling for this. The GF showed a range between 14.9 to 21.1. According to the Fog reading level category, the index achieved from the BIPA-1 instrument is at the level between the grade of a college sophomore and a college junior, while BIPA-2 and BIPA-3 instruments are at the level of a college graduate. Similar to the GF, the SMOG obtained from the analysis are at two different levels. The BIPA-1 instrument is at the 6th-grade level, while the BIPA-2 and BIPA-3 instruments are at the 7th grade. The ARI shows the different level distribution. The BIPA instrument indices are grouped into three grade levels, including 4th grade for BIPA-1, 5th grade for BIPA-3, and 9th grade for BIPA-2.

Analysis of the reading text instrument for the BIPA showed a variation on the score achievement from the cloze test. The average cloze test achievement of the respondents was ranging from 55.2% to 73.3%. Detailed calculation result on the readability indices and the cloze test is presented inFig. .

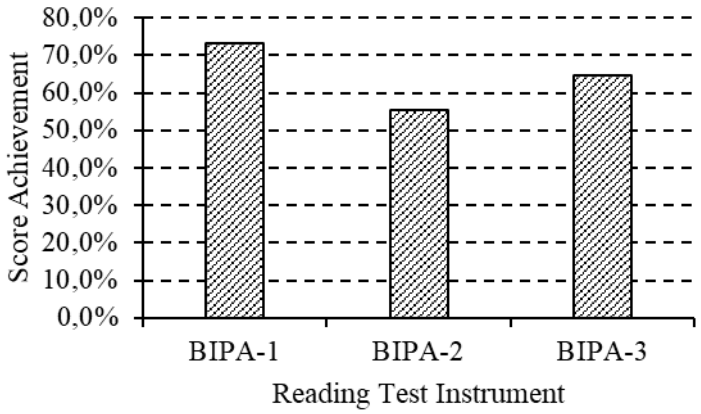

Fig. 1 Score Achievement of Cloze Test for BIPA Reading Text Instrument

The score achievement of BIPA learners as the respondents in this research showed identifiable trends with the readability index, as shown in Table 1. Based on the analysis, the average score obtained from BIPA-1 instrument testing was 3.67 (level of achievement: 73.3%), while for BIPA-2 instrument was 2.76 (level of achievement: 55.2%) and BIPA-3 instrument was 3.24 (level of achievement: 64.8%). The score indicates that the reading test instruments for BIPA-1 and BIPA-3 were at a fairly easy level, while for BIPA-2 was at a moderate level.

The readability of the reading text instrument was not appropriate to the BIPA leveling. The test instrument for the BIPA-2 level was more difficult than that for the BIPA-3 level. In order to assure the learner's knowledge development, appropriate instrument quality leveling is required. Thus, a higher BIPA level should have more difficult test instruments. As a consequence, the test instrument should be revised, or at least exchanged between the BIPA-2 and BIPA-3 instruments.

Data analysis by ANOVA on the readability for cloze test results showed a significant difference on the readability level. A significant difference was obtained from the BIPA-1 and BIPA-2 test scores. The analysis resulted in F value of 2.943 with a probability level of 0.91 ($p < 0.1$). There was an 18.1% of the difference between BIPA-1 and BIPA-2 test achievement. The difference between BIPA-1 and BIPA-3 was only 8.6%, while between BIPA-2 and BIPA-3 was 9.5%.

Correlation analysis between the readability indices and the cloze test achievement showed a significant correlation between the GF and ARI, GF and cloze test achievement, and ARI and cloze test achievement. The correlation of GF to ARI was as much as 99.4% ($p = 0.071$), GFto cloze test achievement as much as -100% ($p = 0.005$), while ARI to cloze test achievement as much as -99.5% ($p = 0.065$). The analysis result showed that there was a negative correlation between GF and ARI to cloze

test achievement. The reversed readibility indexing could cause this for GF and ARI in which a lower index value shows the higher readability.

Even though the grade leveling is not similar among the readability formulas, there was one similar index achievement trend (except for Coleman-Liau Index). The readability grade level obtained from the analysis of the instrument consecutively from the lowest to the highest was BIPA-1, BIPA-3, and BIPA-2. Each formula utilizes a different structural approach. However, each readability formula relies on the number of words and the length of the sentence in the text (Crossley, Allen, & McNamara, 2011).

Readability index analysis also showed that the level up-scaling was not consistent among BIPA grades. Ideally, the increase of test difficulty grade is consistent among the learning stage (Crossley et al., 2011). However, it was not found in this research. Thus, adjustment on the reading test instrument for BIPA needs to be carried out to improve reading comprehension reliability.

Based on the outcome of cloze thest, the examined reading instruments for BIPA-1 and BIPA-3 are categorized as fairly easy, while the BIPA-2 instrument is categorized as moderate. A reading text's readability specifically for a second language learner depends on several factors, including sentence length, the number of known and new vocabulary, and grammatical complexity (Zamanian & Heydari, 2012). Furthermore, some scientist also considers the existence of complex words (with 3+ syllables) as a factor affecting the readability (Loughran & McDonald, 2014). However, the achievement of the cloze test in this research is considered related to the reading texts' length. In summary, the instrument for BIPA-1 only consisted of 39 words, 278 words for the BIPA-2 instrument, and 223 words for the BIPA-3 instrument.

The difference of cloze test achievement was significant between BIPA-1 and BIPA-2 instruments. However, there was no significant difference obtained between the BIPA-1 and BIPA-3 or between BIPA-2 and BIPA-3. This indicates the low discrimination of instrument difficulty. Although there was an increase in average test scores, the instrument still needs further improvement. A test instrument should have a discriminating function to emphasize the difference among grades (Hamada, 2015).

The analysis result indicates that the placement of the test instrument between the BIPA-2 and BIPA-3 was reversed. The examination result showed that the reading test instrument for BIPA-3 was easier than the BIPA-2 instrument. Commonly, test instruments' difficulty should be comparable to the grades (Zamanian & Heydari, 2012). Higher education grades should be provided with more difficult tests to ensure that only those who have obtained the knowledge could complete the test.

Appropriate grade leveling is needed to ensure that the learning development is going on the right track. The assessment of grade level for a particular subject can be determined from the achievement of a specific examination process (Zhang & Misiak, 2015). In this study, the reading text instrument's grade level can be retrieved from the readability index and the cloze test achievement. Thus, the reading text instrument developed for BIPA-2 should be used in BIPA-3 and vice versa.

According to the analysis result, the readability indices assessment for the BIPA reading test instrument was reliable for providing initial information regarding each level's difficulty level. This research implies that GL and ARI's readability index were the most relevant to measure the readability of Indonesian of the BIPA students, shown by the correlation indices. However, this result is contrary to the result of Zamanian, which showed that the Flesch-Kincaid formula was more relevant for second language reading ability measurement (Zamanian & Heydari, 2012: Díaz-Levicoy et al, 2019).

Moreover, even though the other formula's index achievement, such as Flesch-Kincaid Grade Level and SMOG Index, was not significantly correlated with the actual achievement, but both formulas showed a similar trend to the cloze test achievement. Thus, utilizing formula-based indices for initial readability assessment is relevant. This should ease the corpus' placement process to specific audience/grade compatibility (Daud, Hassan, & Aziz, 2013).

Even though the readability indexing formula can assess the readability of the reading test instrument, but actual achievement is the most important indicator. Reading ability is related to the reading's structural aspect and related to the reader aspect, such as their knowledge and interest in the reading text (Wray & Janan, 2013). In developing a test instrument, the test achievement actually acts as the feedback to the developer(Jandaghi, 2011). Thus, the developer should carry out improvements on their test materials according to the result of their instrument examination.

5. Conclusions

The existence of a readability indexing formula helps to evaluate its potential to be used to carry out an initial assessment of the text readability. This is recommended, especially to avoid misplacement of the instrument toward the grade levels. However, pilot testing is also needed to validate the quality of the test instrument. This research proved that the developed reading test instrument for BIPA showed distinct leveling, but there was a misplacement between the reading text of BIPA-2 and BIPA-3 instruments, which has to be fixed. The Gunning Fog Score and Automated Readability Index were the most relevant text readability formula to assess the difficulty level of the test instrument of BIPA.

The development of a qualified and appropriate reading test instrument for BIPA needs to be carried out to achieve its perfection. Moreover, the BIPA program as the media of Indonesian language teaching to the foreign students needs to apply standardized instrument which involves various linguistic aspects, such as grammatical, vocabulary, and complexity within the reading instrument. However, the development of appropriate test instruments is a long continuous process. Thus, further development effort, assessment, evaluation, and improvements should be carried out.

6. Acknowledgement

Acknowledgements are conveyed to Universitas Muhammadiyah Surakarta Research and Community Service Institute that has funded this research.

7. References

Alkhurayyif, Y., & Weir, G. R. S. (2017). Evaluating readability as a factor in information security policies. *International Journal of Trend in Research and Development*, 54–64.

Azizah, R. F., Hs, W., & Lestari, I. (2013). Pembelajaran Bahasa Indonesia bagi Penutur Asing (BIPA) Program CLS (Critical Language Scholarship) di Fakultas Sastra Universitas Negeri Malang tahun 2012. *Vokal, 1*, 1–13.

Binsaleh, S., & Binsaleh, M. (2021). 4P-2E Model: Teaching and Learning Process Through ICT Integration for Private Islamic Schools in Thailand. *Asian Journal of University Education, 16*(4), 71-81.

Cheng, C., & Dunn, M. (2015). Health literacy and the Internet: a study on the readability of Australian online health information. *Australian and New Zealand Journal of Public Health, 39*, 309–314.

Crossley, S. A., Allen, D. B., & McNamara, D. S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language, 23*, 84–102.

Daud, N. M., Hassan, H., & Aziz, N. A. (2013). A corpus-based readability formula for estimate of arabic texts reading difficulty. *World Applied Sciences Journal, 21*, 168–173.

Dunham, B., Yapa, G., & Yu, E. (2015). Calibrating the difficulty of an assessment tool: The blooming of a statistics examination. *Journal of Statistics Education, 23*, 1–33.

Hamada, A. (2015). Linguistic variables determining the difficulty of Eiken reading passages. *Japan Language Testing Association Journal, 18*, 57–77.

Jandaghi, G. (2011). Assessment of validity, reliability and difficulty indices for teacher-built physics exam questions in first year high school. Arts and Social Sciences Journal, 16, 1–4.

Kasule, D. (2011). Textbook readability and ESL learners. *Reading and Writing, 2*, 63–76.

Kouame, J. B. (2010). Using readability tests to improve the accuracy of evaluation documents intended for low-literate participants. *Journal of MultiDisciplinary Evaluation, 6*, 132–139.

Loughran, T., & McDonald, B. (2014). Measuring readability in rinancial disclosures. *The Journal of Finance, 69*, 1643–1671.

Onwuegbuzie, A. J., Mallette, M. H., Hwang, E., & Slate, J. R. (2013). Editorial: Evidence-based guidelines for avoiding poor readability in manuscripts submitted to journals for review for publication. *Research in the Schools, 20*, 1–11.

Rahmat, H., Leng, C. O., & Mashudi, R. (2021). Innovative Educational Practice for Impactful Teaching Strategies through Scaffolding Method. *Asian Journal of University Education, 16*(4), 53-60.

Sabjan, A., Abd Wahab, A., Ahmad, A., Ahmad, R., Hassan, S., & Wahid, J. (2021). MOOC Quality Design Criteria for Programming and Non-Programming Students. *Asian Journal of University Education, 16*(4), 61-70.

Sultan, Rofiuddin, A., Nurhadi, & Priyatni, E. T. (2017). The development of a critical reading learning model to promote university students' critical awareness. *THe New Educational Review, 48*, 76–86.

Tabatabaei, E., & Bagheri, M. S. (2013). Readability of reading comprehension texts in Iranian senior high schools regarding students' background knowledge and interest. *Journal of Language Teaching and Research, 4*, 1028–1035.

Díaz-Levicoy, D., Batanero, C., Arteaga, P., & Gea, M. M. (2019). Chilean Children's Reading Levels of Statistical Graphs. *International Electronic Journal of Mathematics Education, 14*(3), 689-700. https://doi.org/10.29333/iejme/5786.

Cheng, H.-F. (2020). Learning English: A Study of English Novel Reading Camp. *Mediterranean Journal of Social & Behavioral Research, 4*(2), 31-34. https://doi.org/10.30935/mjosbr/9598

Yilmaz, A. (2018). Computers in Reading and Writing Skills through the Motivational Lens: SnagitTM, Screencast and E-mail Services. *Contemporary Educational Technology, 9*(3), 264-283. https://doi.org/10.30935/cet.444110.

Twum, R., Yarkwah, C., & Nkrumah, I. K. (2021). Utilisation of the Internet for Cyberloafing Activities among University Students. *Journal of Digital Educational Technology, 1*(1), ep2101. https://doi.org/10.21601/jdet/10912.

Wray, D., & Janan, D. (2013). Exploring the readability of assessment tasks: The influence of text and reader factors. *Multidisciplinary Journal of Educational Research, 3*, 69–95.

Xia, M., Kochmar, E., & Briscoe, T. (2016). Text readability assessment for second language learners. In Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications (pp. 12–22). San Diego, California: Association for Computational Linguistics.

Zamanian, M., & Heydari, P. (2012). Readability of texts: State of the art. *Theory and Practice in Language Studies, 2*, 43–53.

Zhang, B., & Misiak, J. (2015). Evaluating three grading methods in middle school ccience classrooms. *Journal of Baltic Science Education, 14*, 207–215.