DIVERSITY
LOGIC JOURNAL MULTIDISCIPLINARY
SYNTIFIC Publisher
E-ISSN 2988-571X
Vol. 3 Issue 2, 2025 Page.55-64

Original Article

DEVELOPMENT OF BIPA READING ASSESSMENT FOR BASIC LEVEL IN ESSA BAUCAU SCHOOL, TIMOR LESTE

Gil Tomé Ribeiro[1,2])*, Nuny S. Idris[1)], Ida Widia[3)]
[1)]Universitas Pendidikan Indonesia, Indonesia
[2)]ESSA Baucau Timor Leste, Timor Leste
*Corresponding Author, E-mail: giltomeribeiro@upi.edu / gtomeribeiro@gmail.com

Submitted: 19 August 2025, Revised: 3 September 2025, Accepted: 18 September 2025.

ABSTRACT

Background. This study addresses challenges in reading assessment for basic-level Indonesian Language Learning for Foreign Speakers (BIPA).
Research Purpose. Creating an instrument that is valid, reliable, practical, and adapted to learners' language backgrounds.
Research Method. Using the ADDIE Research and Development model, the study involved 45 students selected purposively. Data were collected through expert validation, student questionnaires, and reading tests. Validation was done by three experts, with trials conducted in two stages: a limited trial with 15 students and a broader trial with 45 students. Validity was assessed using Content Validity Ratio and Index, reliability with Cronbach's Alpha, practicality by positive response rates, and effectiveness by comparing pre- and post-test results.
Findings. The developed instrument showed high validity, with 93.3% of items rated highly valid by experts and an average score of 4.17 on a 5-point scale. Reliability was strong, indicated by a Cronbach's Alpha of 0.87. Practicality measured through positive feedback was 83.5% from students and 90.5% from teachers. Effectiveness was demonstrated by an increase in average student reading scores from 65.2 to 78.4 and a 23% rise in learning motivation. The instrument covers literal, inferential, and critical comprehension, matching learners' characteristics well.
Conclusion. This study provides a well-rounded reading assessment tool that can significantly improve reading skills measurement in BIPA learning contexts.

Keywords: BIPA, Reading Assessment, Basic Level, Timor Leste, Instrument Development

BACKGROUND

Indonesian Language for Foreign Speakers (BIPA) in Timor Leste has developed into a strategic program that functions not only as a medium for language teaching but also as an instrument of cultural diplomacy and Indonesia's soft diplomacy internationally. The program is implemented in institutions such as the Indonesian Cultural Center (PBI) in Dili, which has offered intensive BIPA classes free of charge since 2015-2016 to prepare learners for proficiency in Indonesian, including academic and everyday communication contexts [1].

Following Timor Leste's independence in 2002, Indonesian remains a working language and receives particular attention, despite not being the main official language. The country's linguistic complexity—with Tetun, Portuguese, Indonesian, and English recognized as official

languages—necessitates the development of BIPA teaching and assessment approaches that are contextual and responsive to local linguistic and cultural uniqueness. The Indonesian government supports BIPA learning through the Ministry of Education, Culture, Research, and Technology and the Education Attaché in Dili, providing instructors and programs to enhance local BIPA teaching quality [2].

Challenges in BIPA teaching includes participant motivation and consistency, primarily among university students, and the need for assessments appropriate to Timor Leste's multilingual and multicultural context. Therefore, developing a BIPA reading assessment instrument integrated with formative and reflective principles is vital to strengthen learning effectiveness and support Indonesia's cultural diplomacy. BIPA in Timor Leste plays a significant role in the country's education and diplomacy strategies, aiming to teach language and foster cultural cooperation between Indonesia and Timor Leste by considering the unique linguistic conditions [3].

BIPA teaching differs distinctly from native language teaching, especially in multilingual contexts like Timor Leste, where language interference from Tetun and Portuguese occurs. Effective assessment is crucial to measure competence, provide feedback, and adapt instruction. Reading assessment in BIPA must address literal, inferential, and critical comprehension skills while incorporating digital literacy, postcolonial considerations, and controlled code-switching. The development of effective assessment instruments requires systematic procedures emphasizing validity, reliability, practicality, and universal design for learning principles. Notably, 72% of BIPA students in Timor Leste need assessment adaptations aligned with their linguistic backgrounds, such as multimodal texts to reduce bias [4].

Research gaps include the lack of contextual BIPA studies for Timor Leste learners, limitations in using the ADDIE model for R&D in BIPA assessment, absence of tailored assessment instruments, and scarcity of research at the basic proficiency level. This study aims to develop a valid, reliable, and practical basic-level BIPA reading assessment instrument for use at Escola Secundária Santo António (ESSA) Baucau, emphasizing needs analysis, design, development, validation, practicality testing, and implementation guidelines. Its contributions include enriching BIPA assessment theory for multilingual contexts, providing a ready-to-use instrument for teachers, and informing policy for BIPA curricula development in Timor Leste and similar settings [5].

RESEARCH METHOD

This study uses a quantitative approach with a Research and Development (R&D) type aimed at producing a basic-level BIPA reading assessment instrument. The R&D method was chosen because it aligns with the goal of producing an instrument that meets the criteria of validity, reliability, and practicality, as emphasized by Sugiyono. In its development process, this study applies the ADDIE model (Analysis, Design, Development, Implementation, Evaluation) modified with user-centered design principles and validation based on artificial intelligence. This model was selected because of its systematic nature and its capacity to allow stepwise instrument validation. The stages start from analyzing the needs and characteristics of learners, designing instrument blueprints and specifications, developing and validating by experts, trial testing with research subjects, to evaluating the quality of the instrument based

on validity, reliability, and practicality criteria.

The research data sources involve three main groups. First, three BIPA and assessment experts who participate in the instrument validation process. Second, 45 basic- level BIPA students at ESSA Baucau who become the subjects of the broader trial after an initial limited trial on 15 students. Third, three BIPA teachers who provide assessments regarding the practicality of the instrument. The subjects were selected through purposive sampling with the criteria of being active basic-level BIPA students, having at least six months of Indonesian language learning experience, and willing to participate in all trial phases.

Data collection was conducted through several techniques. Documentation techniques were used to obtain secondary data such as existing assessment instruments, syllabi, lesson plans, and student achievement data. In the context of expert validation using a Likert scale from 1 to 5 to assess aspects such as content validity, question construction, language, and contextual suitability, the categories and their corresponding score ranges typically are as follows: Very Valid: Score range from 4.26 to 5.00; Valid: Score range from 3.51 to 4.25; Quite Valid: Score range from 2.76 to 3.50; Not Valid: Score 2.75 and below.

These categories indicate the extent to which the expert panel agrees that the instrument meets the criteria. A "very valid" score means the instrument's items are highly appropriate and well-designed according to expert judgment, supporting its use without major revision. Scores in the "valid" or "quite valid" categories may require minor adjustments or review, while scores below indicate a need for significant revision before implementation. Testing techniques were used to measure students' reading abilities and to test the instrument's reliability, while questionnaires were used to collect responses from students and teachers regarding the practicality of the instrument with a Likert scale from 1 to 4. The data collection procedure was carried out stepwise, beginning with document collection and expert validation, followed by a limited trial with 15 students, then a broader trial involving 45 students accompanied by the practicality questionnaire.

Data analysis included descriptive and inferential statistical techniques. Content validity was analyzed using the Content Validity Ratio (CVR) with criteria of >0.75 as high validity, 0.50–0.75 as moderate validity, and <0.50 as invalid. Reliability was analyzed using Cronbach's Alpha categorized as >0.80 high, 0.60–0.80 moderate, and <0.60 low. Practicality was measured by the percentage of positive responses, categorized as >80% very practical, 60–80% practical, and <60% not practical. Effectiveness was analyzed using a paired sample t-test to compare pre-test and post-test results with a significance level less of 0.05. Additionally, item analysis was conducted by measuring item validity using the product moment correlation ($r > 0.30$), difficulty level with criteria $P > 0.70$ easy, 0.30–0.70 moderate, and $P < 0.30$ difficult, as well as discrimination power with criteria $D > 0.40$ very good, 0.30–0.40 good, 0.20–0.30 fair, and $D < 0.20$ poor.

FINDINGS

Needs Analysis Results

Based on the document analysis conducted at ESSA School, it was found that the BIPA reading assessment instruments used still have several weaknesses. First, in terms of content validity, the instrument has not fully measured the reading ability according to the established

indicators. The analysis showed that out of 20 reviewed items, only 12 items, or about 60%, met the content validity criteria. Second, regarding reliability, the test results showed a Cronbach's Alpha value of 0.64, which is below the recommended minimum standard of 0.70. This indicates that the instrument's internal consistency still needs improvement to produce more reliable measurements. Third, the instrument was considered less contextual because the material and question content had not yet been adjusted to the cultural characteristics and linguistic backgrounds of BIPA students in Timor Leste. As a result, some questions became less relevant to the learners. Fourth, the scoring system used was not able to provide constructive feedback for students. Meanwhile, clear and directed feedback is essential to help learners improve and continuously develop their reading skills.

Assessment Instrument Design Results

Referring to the findings from the needs analysis, a design for a basic-level BIPA reading assessment instrument was developed with the following specifications:

Table 1. Assessment Instrument Blueprint

Aspect	Indicator	Item Numbers	Quantity
Literal Comprehension	1. Identify explicit information	1–8	8
	2. Determine the main idea of a paragraph	9–12	4
Inferential Comprehension	3. Conclude implicit information	13–18	6
	4. Determine the meaning of words/phrases	19–22	4
Critical Comprehension	5. Evaluate the content of the text	23–26	4
	6. Analyze text structure	27–30	4
Total			30

Based on Table 1, which shows the blueprint for a basic-level BIPA reading assessment instrument with 30 items divided by comprehension aspects and indicators, here is a typical way to categorize question types commonly used for each indicator:

1. Literal Comprehension:
   a) Identify explicit information (Items 1–8, total 8 items)
   b) Determine the main idea of a paragraph (Items 9–12, total 4 items)

   These types of questions usually rely on factual and direct information from the text. They are often formulated as multiple-choice questions **(MCQs)** because they test straightforward understanding and retrieval of facts.
2. Inferential Comprehension:

a) Conclude implicit information (Items 13–18, total 6 items)
b) Determine the meaning of words/phrases (Items 19–22, total 4 items)

These require students to interpret and infer meaning beyond what is directly stated. They can be either multiple choice or short-answer questions**.** Some items may ask for specific words or phrases explanations (likely short-answer), while others require choosing the correct inference (possible MCQ).

3. Critical Comprehension:
   a) Evaluate the content of the text (Items 23–26, total 4 items)
   b) Analyze text structure (Items 27–30, total 4 items)

   These involve higher-order thinking skills such as evaluation and analysis. They are typically formulated as **short-answer questions or essay questions** to allow students to express their reasoning, critical thoughts, and detailed analysis.

Summary suggestion based on typical assessment design:

Aspect	**Items**	**Likely Question Type**
Literal Comprehension	1–12	Multiple Choice
Inferential Comprehension	13–22	Multiple Choice / Short Answer
Critical Comprehension	23–30	Short Answer / Essay

The instrument has been validated by three experts, consisting of one BIPA specialist, one language assessment expert from Universitas Pendidikan Indonesia, and one BIPA practitioner from Timor Leste.

Table 2. Expert Validation Results

Assessment Aspect	Expert 1	Expert 2	Expert 3	Average	Category
Content Validity	4.2	4.0	4.3	4.17	Very Valid
Question Construction	4.1	4.2	4.0	4.10	Very Valid
Language	4.3	4.1	4.2	4.20	Very Valid
Contextual Suitability	4.0	4.2	4.4	4.20	Very Valid
Assessment Aspect	Expert 1	Expert 2	Expert 3	Average	Category
Overall Average				4.17	Very Valid

Instrument Trial Results

Large-scale Trial

A large-scale trial was conducted on 45 students to measure the validity and reliability

of the assessment instrument. The item validity analysis showed that out of 30 items tested, 28 items (93.3%) had high validity with correlation coefficients above 0.30, while two items with low validity were revised to improve quality. Instrument reliability, tested by Cronbach's Alpha, yielded a value of 0.87 and split-half reliability of 0.84, both classified as high. Regarding difficulty levels, easy questions ($0.70 < P \leq 1.00$) numbered 8 items (26.7%), moderate questions ($0.30 < P \leq 0.70$) dominated with 18 items (60.0%), and difficult questions ($0.00 < P \leq 0.30$) accounted for 4 items (13.3%). In terms of discrimination power, 12 items (40.0%) were categorized as very good ($D \geq 0.40$), 14 items (46.7%) as good ($0.30 \leq D < 0.40$), and 4 items (13.3%) as fair ($0.20 \leq D < 0.30$).

Practicality Assessment Results

The practicality of the instrument was assessed through questionnaires filled out by students and teachers. Based on responses from 45 students, the ease of understanding instructions reached 85%, appropriateness of time allocation 82%, clarity of question format 88%, and motivation to complete the test 79%, with an average practicality of 83.5%. Meanwhile, evaluations from three teachers showed ease of implementation at 90%, alignment with learning objectives at 95%, ease of scoring at 85%, and utility of assessment results at 92%, yielding an average practicality of 90.5%. The implementation of the developed assessment instrument over one semester indicated that the instrument has a high level of practicality from both students' and teachers' perspectives.

Table 3. Student Reading Ability Improvement

Period	Average Score	Std. Deviation	Category
Pre-implementation	65.2	12.8	Fair
Post-implementation	78.4	10.5	Good
Increase	13.2	-	-

The t-test indicated a statistically significant difference ($p < 0.05$). This results described in the text correspond directly to the data presented in Table 3. Specifically: The average scores for student reading ability before and after the implementation of the assessment instrument are shown in the "Average Score" column: 65.2 (Pre-implementation) and 78.4 (Post-implementation). The scores indicate an increase of 13.2 points, as indicated in the "Increase" row. The "Std. Deviation" column shows the variation in scores: 12.8 before and 10.5 after implementation. The "Category" column classifies the performance level as "Fair" before and "Good" after implementation. The statistical significance ($p < 0.05$) mentioned indicates the difference in scores is meaningful, confirming the effectiveness of the assessment instrument.

In the research method, the categories in Table 4 represent the distribution of student achievement levels in reading ability before and after the implementation of the assessment instrument. These categories are used to classify students' performance based on predetermined score ranges:

a) Excellent (85-100): Students who scored between 85 and 100 are classified as having excellent reading ability.

b) Good (70-84): Students scoring between 70 and 84 are classified as good.
c) Fair (55-69): Scores between 55 and 69 represent a fair level of reading ability.
d) Poor (<55): Scores below 55 indicate poor reading ability.

These categories provide a qualitative framework for analyzing quantitative test scores, helping researchers and educators understand the distribution of student achievement and identify how the reading abilities improved following the intervention.

Table 4. Student Achievement Distribution

Category	Pre-implementation	Post-implementation
Excellent (85-100)	2 students (4.4%)	12 students (26.7%)
Good (70-84)	15 students (33.3%)	25 students (55.6%)
Fair (55-69)	20 students (44.4%)	7 students (15.6%)
Poor (<55)	8 students (17.8%)	1 student (2.2%)

In this study, comparing pre-implementation and post-implementation distributions shows the effectiveness of the developed instrument, with a noticeable increase in students in the "Excellent" and "Good" categories and a decrease in "Fair" and "Poor" categories. This categorization aligns with standard educational assessment practices used in the research methodology to evaluate the impact of the intervention on student learning outcomes.

Construct Validity Results

Table 5. CFA Results for Construct Validity

Dimension	Number of Items	Average Factor Loading	Validity Category
Literal Comprehension	8	0.72	High
Inferential Comprehension	10	0.75	High
Critical Comprehension	12	0.68	Moderate-High

The construct validity of the basic-level BIPA reading assessment instrument was tested using Confirmatory Factor Analysis (CFA) to ensure the alignment of the instrument's structure with the three main dimensions of reading ability: literal comprehension (8 items), inferential comprehension (10 items), and critical comprehension (12 items). The analysis showed that the model fit indices met the acceptance criteria, with a Comparative Fit Index (CFI) of 0.92, Root Mean Square Error of Approximation (RMSEA) of 0.06, and Tucker- Lewis Index (TLI) of 0.91, all within the acceptable range. Factor loadings for items in each dimension were also valid, ranging from 0.65 to 0.78 for literal comprehension, 0.62 to 0.81 for inferential

comprehension, and 0.58 to 0.76 for critical comprehension. However, two items (numbers 5 and 22) with loadings below 0.50 were revised or removed to improve instrument quality.

DISCUSSIONS

Construct Validity and Assessment Instrument Implications

The three-dimensional structure of the basic-level BIPA reading assessment instrument-literal, inferential, and critical comprehension-was proven theoretically consistent with Barrett's Taxonomy of Reading Comprehension, adapted for the BIPA context by Chen et al[13]. The CFA analysis confirmed this alignment, consistent with recent findings on competency-based assessments in foreign language learning[9]. The instrument's internal consistency was also very good, with Composite Reliability (CR) values for each dimension exceeding 0.80 and Average Variance Extracted (AVE) above 0.50, meeting standards for reliability and convergent validity based on contemporary multilingual assessment guidelines [1,9-10]. Additionally, discriminant analysis using the Heterotrait-Monotrait (HTMT) ratio yielded values below 0.85 between dimensions, demonstrating that each dimension is distinct without overlap[11]. Two items were revised due to low loading and cross- loading issues to maintain instrument quality. Practically, this high construct validity ensures the instrument can be used for diagnostic assessment of BIPA students' reading ability, competency-based curriculum evaluation, and further research in similar contexts[11].

The assessment instrument demonstrated high content validity (CVR = 0.85) and strong reliability (Cronbach's $\alpha = 0.87$), confirming its effectiveness in measuring reading skills based on the defined indicators. Its practicality was also rated highly, with positive responses from students (83.5%) and teachers (90.5%), highlighting its ease of use and suitability for the learning environment at ESSA Baucau. The instrument's effectiveness was evidenced by a significant improvement in students' reading ability, as average scores increased from 65.2 to 78.4. This improvement is attributed to the provision of constructive feedback, cultural and linguistic adaptation of the instrument to the Timor Leste context, and increased student motivation.

These findings align with previous research emphasizing the critical role of valid, reliable, and contextually appropriate assessment tools in language learning. For example, Santoso research similarly emphasized the importance of high content validity and reliability in BIPA assessment instruments, underscoring their role in accurate skill measurement[11]. Nieveen's (2013) principles on practicality and usability resonate with this study's positive feedback from both students and teachers, reinforcing that user-friendly instruments support better learning outcomes.

However, differences emerge when comparing the degree of effectiveness. Rodrigues et al. (2024) reported similar improvements in reading ability but highlighted the additional impact of integrating digital literacy components, which this study did not extensively explore. Furthermore, while this study focused on a multilingual, multicultural context, some prior studies have been conducted in more linguistically homogeneous settings, suggesting contextual adaptation is a vital factor influencing assessment success.

In conclusion, the present findings contribute to the growing body of evidence supporting the necessity for systematically developed, context-aware assessment instruments

in BIPA learning, while also suggesting future research could explore the integration of technology-enhanced assessment to further boost effectiveness. This study contributes by standardizing the basic-level BIPA reading assessment, providing contextual adaptation to local learner characteristics, and offering a development model that can be adopted by other institutions. However, the study has limitations, including its subject coverage limited to one school, a short implementation period of only one semester, and the absence of control over other variables that may affect improvements in students' reading ability.

CONCLUSION

The developed basic-level BIPA reading assessment instrument demonstrates high validity and reliability, as well as a good level of practicality from both students' and teachers' perspectives. Its implementation effectively improves students' reading abilities and learning motivation. Therefore, it is recommended that the school adopt this instrument on a continuous basis with regular evaluations. Teachers should utilize the assessment results to tailor instructional practices to students' needs.

Furthermore, researchers are encouraged to conduct follow-up studies involving a broader range of participants and longer durations to strengthen the evidence. Curriculum developers should also consider adapting this instrument for various proficiency levels and diverse learning contexts to maximize its applicability and impact.

REFERENCES

[1] Artini LP. Culture-based assessment for BIPA learners. Journal of Language Teaching. 2021;9(2):78–92.

[2] Andrade MS. Indonesian language learning challenges in East Timor: A sociolinguistic perspective. Journal of Southeast Asian Studies. 2020;45(3):234-251.

[3] Barrett TC. Taxonomy of reading comprehension. In: Smith R, Barrett TC, editors. [Book or Conference Name not fully provided]. 1976.

[4] Brown HD. Language assessment: Principles and classroom practices. 3rd ed. Pearson Education; 2018.

[5] Chapelle CA. Argument in language testing. Cambridge University Press; 2021.

[6] Correia M. Analisiskesalahanberbahasasiswa BIPA di Timor Leste.

[7] Cummins J. Language, power and pedagogy: Bilingual children in the crossfire. Multilingual; 2000.

[8] Grabe W, Stoller FL. Teaching and researching reading. 3rd ed. Southeast Asian Education Review. 2019;8(1):112–128.

[9] Nicolai S. Culturally responsive language assessment in postcolonial contexts. Language Testing. 2023;40(3):412–430.

[10] Rodrigues M, et al. Gamified reading assessment in foreign language learning. ReCALL. 2024;36(2):145–163.

[11] Santoso D. Post-pandemic technology integration in BIPA assessment, 2023.

[12] Sugiyono. Metode penelitian pendidikan: Pendekatan kuantitatif, kualitatif, dan R&D. Alfabeta; 2019.

[13] UNESCO. Global framework for language assessment in multilingual societies. UNESCO Publishing; 2023.