

RESEARCH ARTICLE

Development of a four-tier gamified diagnostic test to identify students' misconceptions in Ejaan Yang Disempurnakan (Eyd) V

Moh. Faridl Darmawan^{1*} , Faishol Hidayatulloh² , Afif Kholisun Nashoih³ 

¹ Universitas KH. A. Wahab Hasbullah Jombang, Indonesia

² Universitas KH. A. Wahab Hasbullah Jombang, Indonesia

³ Universitas KH. A. Wahab Hasbullah Jombang, Indonesia

***Corresponding Author**

faridldarmawan@unwaha.ac.id

ABSTRACT

In Indonesian language learning, misconceptions regarding the Indonesian Spelling Guidelines (EYD V) are still frequently encountered. These misconceptions indicate underlying conceptual misunderstandings rather than merely technical errors, thereby highlighting the need for a diagnostic evaluation instrument capable of uncovering students' understanding in greater depth. Therefore, this study aimed to develop an evaluation instrument based on the Four-Tier Gamified Diagnostic Test (FT-GDT) to identify students' misconceptions in the Indonesian Spelling Guidelines (EYD V). The instrument consisted of 25 multiple-choice items with five answer options for the first and third tiers, and two confidence options ("confident" and "not confident") for the second and fourth tiers, producing 100 items in total. The development employed a Research and Development (R&D) approach adapted from Borg and Gall. Expert validation indicated high feasibility, with average scores of 86.7% from subject matter experts and 88.3% from media experts. Empirical validity testing showed that most items met the required criteria ($r \geq 0.254$), while reliability testing with 40 students produced a Cronbach's Alpha of 0.87, indicating strong internal consistency. Misconception analysis revealed that students' misconceptions ranged from 17.5% to 45%, with many students expressing high confidence in incorrect answers. These findings demonstrate that the misconceptions were conceptual rather than merely technical errors. Overall, the FT-GDT instrument proved effective in identifying misconceptions in EYD V learning, highlighting its potential as a diagnostic and formative evaluation tool in higher education.

Keywords: Four tier diagnostic test; evaluation; EYD V.

ARTICLE HISTORY

Received: September 28, 2025

Accepted: July 3, 2026

Published: July 27, 2026

LICENSE

Copyright © 2026 The Author(s). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY 4.0) License.

<https://creativecommons.org/licenses/by/4.0/>.

How to Cite This Article

Moh. Faridl Darmawan, Faishol Hidayatulloh, & Afif Kholisun Nashoih

.2026. Development of a four-tier gamified diagnostic test to identify students' misconceptions in Ejaan Yang Disempurnakan (Eyd) V. *Ghancaran: Jurnal Pendidikan Bahasa dan Sastra Indonesia*, Volume 8(1), 379—407. <https://doi.org/10.19105/ghacaran.v8i1.22080>

INTRODUCTION

Indonesian occupies a central role in the academic and professional life of Indonesian citizens, serving as the primary medium of instruction, written communication, and formal expression across all levels of formal education. Proficiency in written Indonesian is contingent upon the mastery of established orthographic conventions, among which spelling stands as a foundational element of academic literacy. In Indonesia, the official reference for spelling norms is *Ejaan yang Disempurnakan* (EYD), first introduced in 1972 and subsequently revised to reflect evolving linguistic usage. The most recent version, EYD Edisi V (EYD V), was issued in 2022 by the Badan Pengembangan dan Pembinaan Bahasa, superseding the previously used *Pedoman Umum Ejaan Bahasa Indonesia* (PUEBI). EYD V introduced refinements in capitalization, punctuation, loanword adaptation, and abbreviation, making it a critical competency domain for higher education students. Against this backdrop, ensuring that students not merely memorize but genuinely understand EYD V conventions becomes an urgent priority for Indonesian language educators.

Learning evaluation is an essential component in the modern education process, not only to measure students' cognitive understanding but also as a diagnostic tool for misconceptions (Rahmawati et al., 2022; Ramly & Abduh, 2018; Ratnasari & Mubasiroh, 2022). Amidst the development of technology and digital-based education, the need for adaptive, accurate, and interactive evaluative instruments is increasingly urgent (Farida et al., 2020). One of the major challenges in learning Indonesian in higher education is students' low understanding of the rules of the General Guidelines for Indonesian Spelling, now formally designated as *Ejaan yang Disempurnakan edisi V* (EYD V), which form the foundation of academic and professional literacy (Febrina & Yusuf, 2021; Setiadin et al., 2024).

Empirical evidence substantiates this concern. A study by Khairani et al. (2025) examining formal media publications in Medan found that errors in capitalization, punctuation, and loanword adaptation were among the most pervasive violations of EYD V norms, indicating that such errors are not confined to novice writers but extend to educated adults in professional settings. In higher education specifically, Zega et al. (2024) documented persistent spelling errors in academic journal articles authored by university students of an Indonesian Language Education programme, with compound word writing, hyphenation, and punctuation among the most frequently misapplied rules. Furthermore, Putri and Handayani (2024) reported that a substantial proportion of students at STKIP PGRI Pacitan demonstrated limited awareness of the conceptual changes introduced in EYD

V compared to its predecessor PUEBI, suggesting that the policy transition has not been matched by a corresponding shift in students' conceptual understanding. Taken together, these findings indicate that EYD V errors are not incidental but reflect systematic conceptual misconceptions, a pattern that cannot be effectively diagnosed or remediated through conventional assessment instruments alone.

The phenomenon of misconceptions in the field of language is not unique to Indonesia. Globally, language misconceptions are a major obstacle to academic communication competence (Nazriani & Arsad, 2020). Various international studies have shown that foreign language learners in different countries still experience conceptual misunderstandings related to orthographic systems and spelling conventions in the target language. The study *International Perspectives on Spelling and Writing in Different Orthographies* (Joshi et al., 2022) revealed that orthographic and spelling errors are universal problems found across various language systems and are not limited to English alone. In addition, the study *Is English Spelling Chaotic? Misconceptions Concerning Its Irregularity* (KESSLER, 2003) demonstrated that many learners possess misconceptions regarding English spelling systems, particularly the assumption that English orthography is entirely irregular and inconsistent. Another study entitled *Orthographic Differences and English L2 Spelling Errors* (Park, 2012) also found that learners of English as a second language experienced conceptual misconceptions in spelling due to differences between the orthographic systems of their first language and the target language.

Similar phenomena are also evident in the context of Indonesian language learning, as reflected in the persistent errors in the use of capital letters, punctuation marks, and loanword spelling in formal academic contexts (Khairani et al., 2025; Rahman et al., 2017; Zega et al., 2024). This condition indicates a fundamental conceptual misconception and is not merely a technical error. Therefore, a diagnostic evaluation instrument is needed that is capable of deeply revealing the flawed mindset behind the linguistic errors of learners (Javidanmehr & Sarab, 2017).

This condition is particularly ironic, considering that students are the intellectual future and citizens of Indonesia who should have a good command of their national language (Sugilastini, 2022; Wirahyuni, 2022). A lack of mastery of correct spelling rules reflects a weak foundation in language literacy, which should be a key asset in the academic world and professional life (Abdurrasyid, 2024; Nabila et al., 2024; Putri & Handayani, 2024).

This problem indicates the need for systematic improvement efforts, not only in the technical aspects of learning but also in the evaluation stage. As such, this research not only

contributes to the development of learning evaluation innovations but also makes an important contribution to the planning and development of the Indonesian language course curriculum. With the Four-Tier Gamified Diagnostic Test-based evaluation instrument, lecturers and program administrators can obtain a more accurate picture of the gaps in student understanding, enabling them to design more targeted and data-driven learning strategies (Caleon & Subramaniam, 2010).

By contrast, the evaluation methods used so to date remain conventional, limited to multiple choice or essay questions, which can only measure factual knowledge without attending the aspects of belief or reasoning behind the answers given (Jumadi et al., 2023). This type of evaluation has not been able to reveal sufficient in depth the location and type of misconceptions experienced by students (Widiyanto et al., 2018). In view of these problems, innovation is needed in the form of an evaluation that is able to explore understanding more comprehensively and motivate students to actively engage in the learning process, as well as identify the level of understanding and misconceptions of students (Pahlefi, 2022; Tomasouw & Marantika, 2023). This evaluation is conducted to collect empirical data on learners' ability to understand language rules and implement them in verbal communication (Lin et al., 2024; Shahmirzadi, 2023).

Several studies have examined the effectiveness of the Four-Tier Diagnostic Test (FTDT) in identifying misconceptions across academic disciplines. Caleon and Subramaniam (2010) conducted a foundational study that established the diagnostic architecture of the four-tier format by combining answer selection, confidence rating, reason selection, and confidence in reasoning. Their findings demonstrated that students frequently hold misconceptions with high confidence and provided the methodological basis for subsequent diagnostic instrument development across disciplines.

In the sciences, Istiyono et al. (2023) developed a Four-Tier Diagnostic Test in physics education using modern test theory, confirming the instrument's superiority over traditional formats in classifying students' conceptual errors. Diani et al. (2019) applied a similar approach to fluid mechanics, finding that a substantial proportion of students held misconceptions with high confidence, thereby reinforcing the conceptual distinction between knowledge gaps and genuine misunderstanding. More recently, Aristiawan et al. (2025) extended this diagnostic model to algebraic thinking in middle school mathematics, demonstrating its cross-disciplinary applicability.

Within the domain of linguistics, the application of the four-tier diagnostic approach remains limited. Laila and Nashoih (2024) developed a four-tier diagnostic instrument for

balaghah (Arabic rhetoric) material using the Wordwall platform, demonstrating that gamification can be integrated into diagnostic assessment without compromising diagnostic accuracy. Their study provided the most proximate precedent for the present research, establishing both the feasibility of the four-tier framework in a linguistic context and the utility of Wordwall as a delivery platform. However, their instrument was developed for Arabic rhetoric, a markedly different linguistic domain, and was not designed to address the specific orthographic conventions governing Indonesian spelling under EYD V. The existing research in Indonesian language evaluation has focused primarily on measuring outcomes using conventional instruments (Rahmawati et al., 2022; Ratnasari & Mubasiroh, 2022), leaving the diagnostic dimension of spelling misconceptions largely unexplored.

Based on this review, a clear research gap is identified: while the four-tier diagnostic test has been widely validated in the sciences and applied in select areas of language learning, its use in the context of Indonesian spelling instruction, particularly under the updated EYD V framework, has not previously been documented. Furthermore, the integration of gamification elements into a four-tier diagnostic instrument for Indonesian language material represents an uncharted methodological space. This study directly addresses this gap by developing an accessible, gamified Four-Tier Diagnostic Test (FT-GDT) specifically designed to identify misconceptions in EYD V material.

The novelty of this research is articulated on two distinct dimensions. Methodologically, this study is the first to extend the four-tier diagnostic assessment paradigm from the natural sciences and select areas of language learning to the domain of Indonesian spelling instruction within the EYD V framework. This contribution fills a critical void in Indonesian language education research by demonstrating that tiered diagnostic assessment is both applicable and effective in capturing the nature and degree of linguistic misconceptions. Pedagogically, the integration of gamification through the Wordwall platform introduces an innovation in assessment design that simultaneously maintains diagnostic rigour and enhances learner engagement, a combination not previously documented in Indonesian language education research. Together, these contributions offer both a validated measurement tool and a replicable methodological approach for future instrument development in language education contexts.

This study fills this gap by developing an Android-based Four-Tier Gamified Diagnostic Test that integrates four levels of analysis, namely answers, confidence, reasons, and confidence levels (Jannah et al., 2022; Salele et al., 2025), with a gamification approach to increase student engagement. The advantage of this model lies in its ability to detect

misconceptions more deeply and provide feedback that can be used directly by lecturers in designing learning interventions.

This research is reinforced by the digital transformation in higher education, which demands mobile-friendly and interactive evaluation instruments. In the context of the Merdeka Belajar curriculum, such evaluation instruments can also support continuous formative assessment based on the analysis of students' learning needs. Thus, this research makes a significant contribution to strengthening the Indonesian language course curriculum. The results of this development are expected to serve as a tool for validating concepts and mapping students' weaknesses.

METHOD

This study is classified as development research, employing a Research and Development (R&D) approach based on the model adapted from Borg and Gall. Development research was chosen because the primary objective is to produce, validate, and refine an educational product, specifically a Four-Tier Gamified Diagnostic Test (FT-GDT) instrument, rather than to test hypotheses about existing phenomena. This approach involves a systematic and iterative process of design, expert validation, empirical testing, and revision, with the ultimate goal of producing a product that meets established quality and usability standards (Welter. R. & Meredith. D, 2003).



Figure 1. Nine Stages of the Adapted Borg and Gall Research Procedure

The research procedure followed ten stages proposed by Borg and Gall, adapted to the scope and resource constraints of this study. As illustrated in Figure 1 above, the adapted stages encompassed: (1) Research and Information Collecting: needs analysis through classroom observation and curriculum review; (2) Planning: defining learning objectives, instrument design, and item specification; (3) Develop Preliminary Form of Product: initial construction of the 25-item FT-GDT instrument on the Wordwall platform; (4) Preliminary Field Testing: expert validation by subject matter and media experts; (5) Main Product Revision: instrument revision based on expert feedback; (6) Main Field Testing: limited empirical trial with 30 students for validity and reliability testing; (7) Operational Product Revision: revision based on trial results; (8) Operational Field Testing: misconception analysis with 40 students and usability evaluation with 60 students; (9) Final Product Revision: refinement based on operational testing outcomes. Due to resource and time constraints, the study was conducted through Stage 9.

The research involved multiple participant groups. For expert validation, two subject matter experts, lecturers specializing in Indonesian language and linguistics, and two media experts with backgrounds in educational technology assessed the instrument's feasibility.

For the limited trial and reliability testing, 30 students enrolled in an Indonesian language course at a higher education institution participated. For the misconception analysis, a separate group of 40 students from the same program completed the full FT-GDT instrument. Additionally, 60 students responded to the student experience questionnaire to evaluate the instrument's usability and interactivity. All student participants were purposively selected based on their enrollment in courses covering EYD V material.

The data collection instruments included three types, namely a validator questionnaire, a student response questionnaire, and a learning evaluation test instrument. The validator questionnaire was employed to obtain an assessment of the instrument's feasibility in terms of content, language, appearance, and technical presentation of gamification. The student response questionnaire was employed to determine users' perceptions of the instrument's ease of use and usefulness. In addition, the evaluation test was employed to measure student learning outcomes and detect any misconceptions.

Data analysis techniques are carried out in several stages. First, expert validation and student response analysis were conducted by calculating the average score using a Likert scale to assess the feasibility and acceptability of the product (Astuti et al., 2023). Second, item validity was tested using the Pearson Product Moment technique and compared with the r -table value at a certain significance level (Kirbulut & Geban, 2014). Third, the reliability of the instrument was tested using Cronbach's Alpha formula to determine the internal consistency of the items (Cirocki et al., 2025; Prabowo, Herman, Fatimah, et al., 2023). Fourth, the analysis of the level of difficulty of the questions was carried out by calculating the proportion of students who answered correctly on each question item. The question items were categorized as easy ($P > 0.7$), medium ($0.3 < P \leq 0.7$), and difficult ($P \leq 0.3$) (Prabowo, Herman, & Fatimah, 2023). Fifth, discrimination analysis is used to determine the ability of items to distinguish between participants with high and low abilities by calculating the difference between the proportion of correct answers from the top and bottom groups relative to the total number of participants (Negoro & Karina, 2019).

RESULTS AND DISCUSSION

The development of the Four-Tier Gamified Diagnostic Test on the EYD V material as a tool to identify students' misconceptions in Indonesian language courses were carried out through a series of systematic R&D stages, starting from needs analysis to limited trials. This process was carried out to answer the problem formulation regarding how the instrument was developed so that it could function optimally in detecting students' misconceptions.

The initial stage began with curriculum analysis and observation of the learning process taking place in the classroom. Researchers found that EYD V material, which covers spelling, punctuation, word writing, and loanwords, is still taught conventionally and is not very interesting for students. Observations and in-depth interviews showed that many students were confused in distinguishing between similar spelling rules, such as the use of capital letters in certain contexts, compound word writing, and the use of punctuation according to the rules. This finding was corroborated by the results of an exploratory questionnaire, which showed that most students felt low confidence when answering language-related questions and frequently relied on guessing rather than conceptual understanding (Widiyanto et al., 2018).

Based on these findings, the researchers designed a four-tier evaluation instrument that not only measures correct or incorrect answers but also explores students' confidence and reasons for choosing their answers. The four levels consist of: (1) choice of answer to the concept question, (2) level of confidence in the answer, (3) logical reasoning behind the choice of answer, and (4) confidence in the correctness of the reasoning given (Aksoy & Erten, 2022; Fratiwi et al., 2017; Yang & Lin, 2015). With this format, the instrument is able to identify not only conceptual errors, but also the level of misconceptions and students' confidence in their understanding.

The developed product presents an evaluation instrument in the form of 25 multiple-choice questions with 5 options (A, B, C, D, and E) for questions at the first and third levels, and the options "confident" and "not confident" for questions at the second and fourth levels. The questions were compiled based on 15 linguistic aspects in EYD V, as shown in the table below.

No.	Basic Competency	Question Indicators	Main Material	Cognitive Level	Question No.
1	Understanding the use of capital letters	Determining the correct use of capital letters in sentences	Capital letters	C1 (Recall)	1, 2
2	Applying the rules for punctuation	Identifying the use of commas and colons in sentences	Punctuation	C2 (Understanding)	3, 4, 5
3	Writing prepositions correctly	Determining the correct spelling of the prepositions "di" and "ke"	Prepositions (di/ke)	C2 (Understanding)	6, 7
4	Writing compound words according to the rules	Determining the spelling of compound words	Compound words	C3 (Apply)	8, 9

5	Using hyphens in writing	Determining the use of hyphens in repeated words and compound forms	Hyphen	C2 (Understanding)	10, 11
6	Writing repeated words correctly	Identifying appropriate word repetition	Repeated words	C1 (Remembering)	12
7	Determining the spelling of loanwords	Identifying the form of foreign words that are absorbed according to the rules	Loanwords	C2 (Understanding)	13, 14
8	Writing the correct hyphenation	Determine how to hyphenate words at the end of a line according to the rules	Word division	C3 (Apply)	15
9	Determining the correct way to write numbers	Choosing the form of writing numbers and numerals in sentences	Writing numbers	C2 (Understanding)	16, 17
10	Writing abbreviations and acronyms correctly	Determining the writing of abbreviations and acronyms according to EYD V	Abbreviations and acronyms	C2 (Understanding)	18, 19
11	Using italics according to function	Determining the use of italics for titles and foreign terms	Italics	C2 (Understanding)	20
12	Determining the use of quotation marks	Determining quotation marks in direct quotations or titles of works	Quotation marks	C3 (Apply)	21, 22
13	Determining the writing of passive verbs	Determining the appropriate form of affixed passive verbs	The prefix di- (passive verbs)	C2 (Understanding)	23
14	Determining the form of cooperation in writing letters	Determining the correct form of the word "cooperation" in a formal context	Formal compound words	C3 (Apply)	24
15	Identifying errors in sentences	Analyzing sentences containing spelling errors	Common spelling errors	C4 (Analyzing)	25

Table 1. Evaluation instrument grid for EYD V material

Question development refers to the cognitive taxonomy starting from the levels of remembering (C1), understanding (C2), applying (C3), to analyzing (C4) (Anderson & Krathwohl, 2001; Cressa & Adni, 2023). For example, the competency of understanding the use of capital letters is measured through two questions that require students to remember the rules. In addition, other aspects such as compound word writing or spelling error analysis require higher-order thinking skills, including application and analysis in the context of sentences. The figure below illustrates how the questions are developed according to the four-level diagnostic approach.

SOAL 2

Tier 1:
 Penulisan yang benar sesuai kaidah huruf kapital dalam PUEBI adalah:
 A. Presiden republik Indonesia akan mengunjungi Sumatera Barat.
 B. Presiden Republik Indonesia akan mengunjungi sumatera barat.
 C. presiden Republik Indonesia akan mengunjungi Sumatera Barat.
 D. Presiden Republik Indonesia akan mengunjungi Sumatera Barat.
 E. Presiden Republik indonesia akan mengunjungi Sumatera barat.

Tier 2:
 Seberapa yakin Anda dengan jawaban tersebut?
 A. Yakin
 B. Tidak yakin

Tier 3:
 Apa alasan Anda memilih jawaban tersebut?
 A. Huruf kapital digunakan untuk nama jabatan jika merujuk pada orang tertentu. dan nama wilayah.
 B. Huruf kapital hanya digunakan untuk nama wilayah.
 C. Kata "presiden" tidak perlu huruf kapital meskipun merujuk pada orang tertentu.
 D. Semua kata dalam kalimat resmi harus menggunakan huruf kapital.
 E. Kata depan dalam nama wilayah tidak perlu kapital.

Tier 4:
 Seberapa yakin Anda terhadap alasan yang Anda pilih?
 A. Yakin
 B. Tidak yakin

Figure 1. Display of question formats with four levels

Based on Figure 1 above, this product consists of four layers of questions (four-tier) for each item (Fariyani et al., 2015), namely:

1. First Tier – Conceptual Questions: Students are asked to choose the answer they think is most appropriate in relation to a particular EYD V concept, such as the use of punctuation or the spelling of loanwords.
2. Second Tier – Level of Confidence in the Answer: Students determine their level of confidence in the answer they have chosen, with options such as "confident" or "not confident".
3. Third Tier – Reasons or Justifications: Students choose the reasons or logical reasoning behind their answers. These reasons have been compiled based on various possible common misconceptions.
4. Fourth Tier – Confidence in Reasons: Students are again asked to assess their level of confidence in the truth of the reasons they have chosen.



Figure 2. Initial display of the instrument.

Figure 3. Question displays at the first level

The initial appearance of the instrument was developed and designed with a simple visual approach. This simplicity aims to reduce the cognitive load on students in the early

stages and facilitate their focus on the evaluative content presented (A. F, 2018). The use of limited visual elements allows learners to immediately understand the workflow without being distracted by irrelevant aesthetic aspects. In the context of digital learning, the simplicity of this display is an added value that supports the principles of accessibility and ease of use across devices, whether through computers or gadgets.

However, the development results revealed limitations in the Wordwall platform, specifically the absence of a dedicated space to explicitly insert instructions before the test begins. This condition poses a challenge, given that instructions play an important role in tiered evaluation instruments such as the Four Tier Gamified Diagnostic Test (FTGDT). Without adequate instructions, students risk misunderstanding the four-level format, which includes answers, confidence, reasons, and confidence in reasons (Maison et al., 2021). To overcome this obstacle, the researchers intervened by providing separate instructions before the test was administered so that participants would have a complete understanding of how to complete the instrument systematically and in accordance with the established evaluation design.

Figure 3 shows the question page at the first level in the EYD V evaluation instrument, which focuses on basic spelling aspects, such as the use of capital letters, punctuation, and loanwords. The questions are arranged in a multiple-choice format with five answer options, consisting of one correct answer and four distractors that are designed systematically. The distractors are not arranged randomly, but based on common types of mistakes often made by students, so that they can reveal the conceptual understanding or misconceptions that test takers may have (Listyowati, 2025).



Figure 4. Display of the second-level question page. **Figure 5.** Third-level question page display

Figure 4 shows the second-tier page, which measures students' metacognitive awareness by prompting them to reflect on their confidence in their first-tier answer. The interface maintains simplicity, presenting only two clearly labeled options: "confident" and

"not confident." This metacognitive component is theoretically significant, as it helps distinguish students who understand the concept from those who merely guessed correctly (Yang & Lin, 2015).

Figure 5 displays the third-tier page, where students select the reasoning underlying their first-tier answer from five options. These options were designed to represent a range of plausible explanations, including theoretically grounded rationales and common misconceptions. This tier is essential for diagnostically differentiating between students who answered correctly for the right reasons and those who may have arrived at the correct answer through faulty reasoning (Fратиwi et al., 2017). Figure 6 shows the fourth-tier page.



Figure 6. Display of the fourth level question page.

Figure 6 shows the fourth-tier confidence page, which mirrors the second tier but applies to the students' chosen reasoning. Together, the data from all four tiers allow for a comprehensive classification of each student's response as reflecting conceptual understanding, partial understanding, misconception (incorrect answer + high confidence), or a knowledge gap (incorrect answer + low confidence) (Salele et al., 2025). This four-category diagnostic framework is one of the key strengths of the FT-GDT over conventional assessment formats.

Validity of the EYD V Evaluation Instrument

The validation of the EYD V evaluation instrument was conducted to ensure that each item developed in the Four-Tier Gamified Diagnostic Test format met both substantive and technical eligibility criteria. The validation process involved assessment by subject matter experts with expertise in Indonesian linguistics and language teaching in higher education, focusing on four main aspects: (1) suitability of material with learning outcomes, (2) depth and scope of material, (3) clarity of language and question formulation, and (4) conformity with EYD V rules and learning objectives (Laila & Nashoih, 2024).

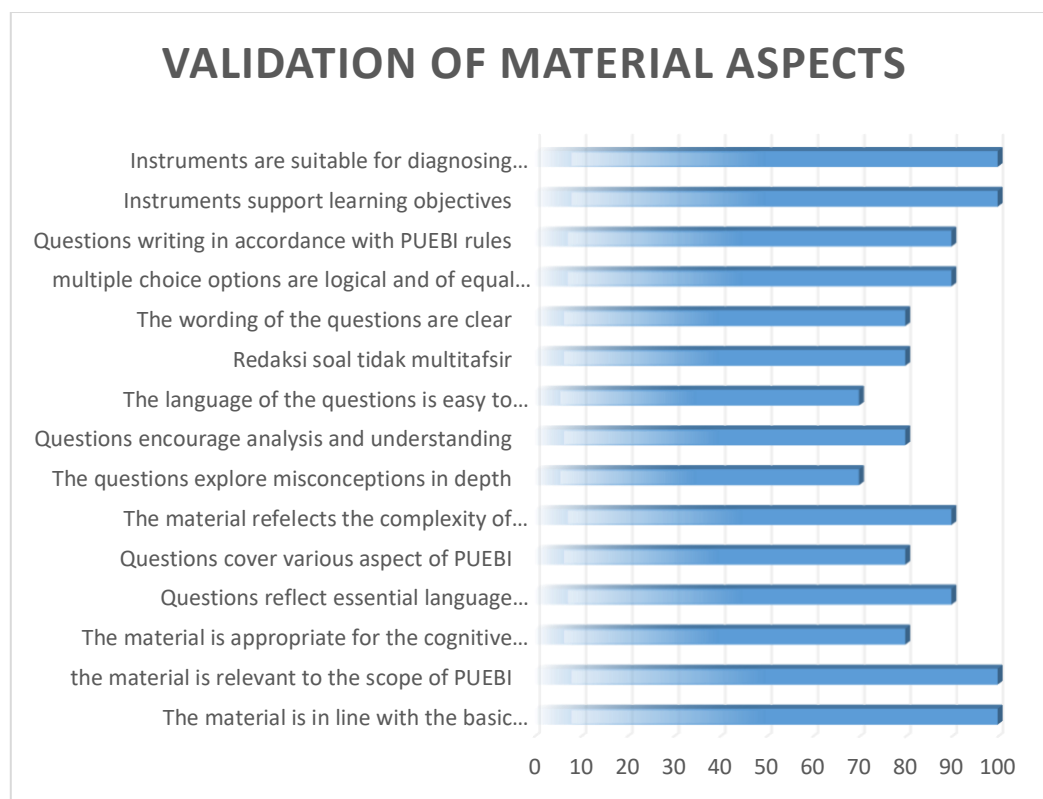


Figure 7. Validation of material aspect

The aspect of material suitability with learning outcomes covers indicators of alignment with basic competencies, scope of EYD V, cognitive level appropriateness, and relevance to essential language competencies. Validation results showed that two indicators, material alignment with basic competencies and material relevance to the scope of EYD V, received maximum scores (100%) from both experts, while material suitability with students' cognitive level received 80% and the indicator of questions reflecting essential language competencies received 90%. These results confirm the strong alignment between the instrument content and the intended learning outcomes (Anderson & Krathwohl, 2001).

Regarding the depth and scope of material, the indicators of questions covering various aspects of PUEBI EYD V and questions encouraging analysis and understanding each scored 80%, while the indicator of material reflecting linguistic complexity scored 90%. The indicator of questions that explore misconceptions in depth received the lowest score (70%), indicating that the instrument's ability to probe specific and deep misconceptions could be further strengthened in subsequent revisions (Prabowo, Herman, Fatimah, et al., 2023).

In terms of language clarity and question formulation, three indicators (unambiguous question wording, clear instructions, and questions that encourage critical thinking) each scored 80%. The indicator of question language being easy to understand received 70%,

suggesting that the language of some items may require simplification to avoid potential misinterpretation. The indicator of answer choices being logically organized and of equal difficulty received a high score (90%), reflecting the quality of the distractor design.

The aspect of conformity with PUEBI EYD V rules and learning objectives received the highest scores overall. The indicator of question writing in accordance with EYD V rules scored 90%, while the indicators of instrument supporting the achievement of learning objectives and instrument suitability for diagnosing misconceptions both received maximum scores (100%). These results confirm the instrument's strong functional and diagnostic validity.

Overall, Expert 1 awarded an average score of 85.3% and Expert 2 an average of 88%, yielding a combined average of 86.7%. This figure meets the criteria for high validity, indicating that the instrument's content, scope, language, and structure are well-aligned with the learning outcomes and cognitive characteristics of higher education students. These findings are consistent with the validation standards applied in comparable instrument development studies (Aristiawan et al., 2025).

Trial of the EYD V Evaluation Instrument based on the Four Tier Gamified Diagnostic Test

The Four-Tier Gamified Diagnostic Test on the Indonesian Spelling Guidelines (EYD V) material was conducted through a limited trial involving a number of students in Indonesian language courses. The purpose of this trial was to assess the instrument's effectiveness in mapping students' misconceptions in a comprehensive manner, and to evaluate users' experiences with the gamified platform (Rawh et al., 2020).

Based on the trial results, the implementation of the application showed that this instrument was effective in mapping the various types of misconceptions experienced by students. Each question in the application provided an overview not only of the correct or incorrect answers, but also of the students' level of confidence in their answers and the arguments underlying their choices. Thus, this application enabled the detection of misconceptions based on a combination of answers, reasons, and students' level of confidence (Laila & Nashoih, 2024).

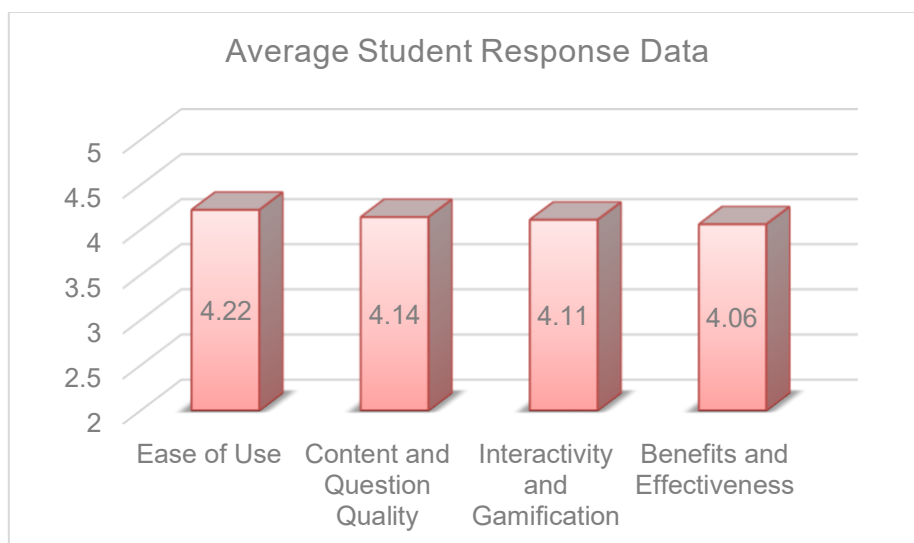


Figure 8. Average data acquisition of student responses on each aspect

Based on the results of a questionnaire given to 60 students, the response to the EYD V evaluation instrument showed very positive results. The analysis was conducted on four main aspects, namely ease of use, content and question quality, interactivity and gamification, and the benefits and effectiveness of the instrument in detecting misconceptions.

In terms of ease of use, the majority of students gave high scores for the accessibility and interface design of the application. This shows that the developed application has fulfilled the principles of usability, namely that it is easy to use, intuitive, and does not require special training. Furthermore, the quality of content and questions also received good responses. Students assessed that the questions presented were in line with the EYD V material they had studied and were able to explore their understanding in greater depth. This indicates that the instrument design not only tested surface-level understanding but also successfully traced conceptual errors that might not be detected through conventional evaluation methods (Jumadi et al., 2023).

In terms of interactivity, most respondents appreciated interactive features such as automatic scoring, immediate feedback on answers, and gamification elements. Students stated that these features made the experience of working on questions more enjoyable and interesting (Latifah & Damayanti, 2022). In fact, many of them felt that the automatic feedback helped them understand their mistakes. The average score for this aspect was also 4.11, indicating good acceptance, so that the use of this application is not only cognitively beneficial but also affectively, creating a positive learning atmosphere.

Finally, the aspect of benefits and effectiveness received an average score of 4.06, with all indicators falling into the "Good" category. This indicates that students consider this instrument to be quite effective in supporting concept understanding and identifying misconceptions (Rawh et al., 2020). Students feel that they have gained a better understanding of EYD V material after using this instrument. They stated that using the instrument made them understand EYD V material better and feel more confident in answering language questions.

Validity and Reliability of the Evaluation Instrument

Validity and reliability testing are essential components in instrument development to ensure that the tool produces accurate and consistent data across contexts and administrations (Sari & Sunyono, 2019). Validity refers to the degree to which an instrument measures what it is intended to measure, while reliability reflects the instrument's capacity to produce stable and consistent results upon repeated use under comparable conditions (Aristiawan et al., 2025).

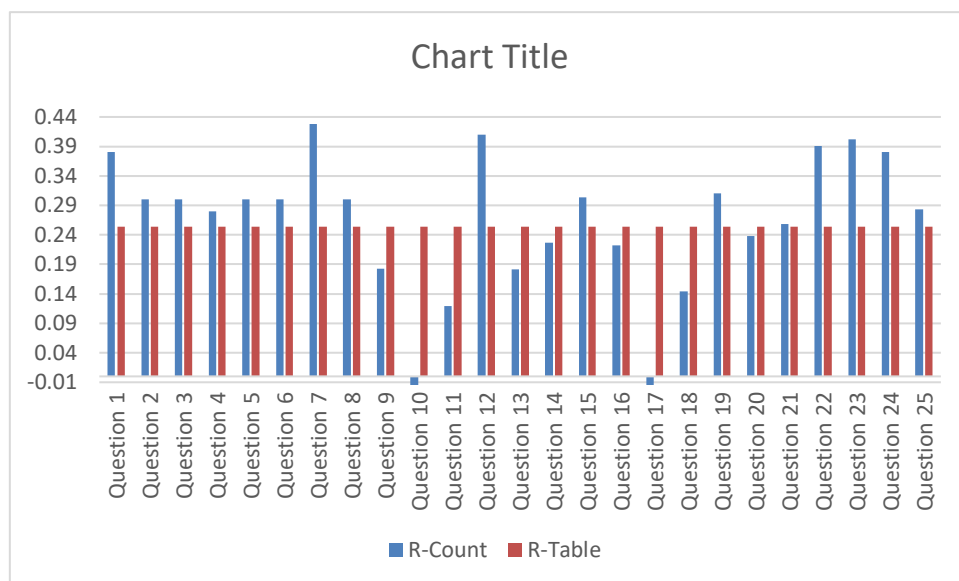


Figure 9. R-Calculated and R-Table correlation values as determinants of the validity of the evaluation instrument

Based on the validity calculations, 25 items were tested using item-total correlation. Each item's correlation coefficient (R_Count) was compared against the R_Table value of 0.254, which corresponds to a sample size of 60 at a significance level of $\alpha = 0.05$. Items with $R_Count \geq 0.254$ were classified as valid. From this analysis, 16 of the 25 items had correlation values meeting or exceeding the r-table threshold, and were therefore classified

as valid. These valid items include Questions 1, 2, 3, 4, 5, 6, 7, 8, 12, 15, 19, 21, 22, 23, 24, and 25. This shows that most items in the instrument have met the validity criteria and are suitable for use in the measurement or evaluation process.

Conversely, the remaining nine items had R_{Count} values below R_{Table} and were classified as invalid. Some items, notably items 10 and 17, yielded negative correlation values, indicating that performance on these items was inversely related to overall test performance. This pattern suggests that these items are not only poor measures of the intended construct but may actively reduce the instrument's overall reliability if retained without revision (Kirbulut & Geban, 2014). These findings underscore the need for systematic item revision prior to broader implementation.

Explanation	Value
Number of Items	25
Number of Respondents	30
Total Item Variance	5.87
Total Score Variance	23.69
Cronbach's Alpha	0.79

Table 2. Reliability Analysis Results

Reliability was assessed using Cronbach's Alpha, which measures the internal consistency of items, that is the degree to which all items collectively measure the same construct in a stable manner. Based on the reliability analysis of the 25 multiple-choice items with 30 respondents, the Cronbach's Alpha value obtained was 0.79, as shown in Table 2. This value falls within the acceptable range for diagnostic instruments and indicates adequate internal consistency (Cirocki et al., 2025). While a higher alpha would be desirable for high-stakes assessment, the current value is appropriate for a formative diagnostic instrument at this stage of development and is consistent with values reported in comparable studies (Prabowo, Herman, & Fatimah, 2023).

Level of Difficulty and Discrimination Power

The difficulty level was calculated based on the ratio between the number of students who answered correctly in the first tier and the total number of test participants. This ratio represents the proportion of students who were able to answer correctly, thus providing an overview of the degree of difficulty of each question in revealing students' understanding of EYD V rules. Questions with a moderate level of difficulty are considered most ideal in a diagnostic context, as they can show differences in the level of understanding among students more accurately. Within the FTDT framework, this aspect of difficulty is important for identifying misconceptions, ignorance, or accurate understanding of the linguistic concepts being tested (Istiyono et al., 2023).

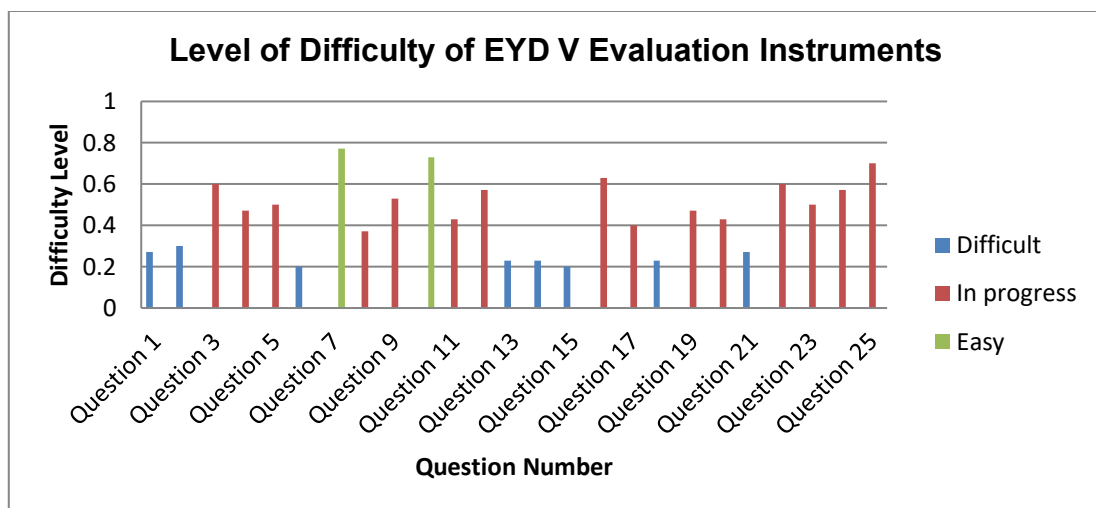


Figure 10. Results of difficulty level analysis

The difficulty analysis of the 25 items revealed a fairly balanced distribution, with a predominance of medium-difficulty items. Of the 25 items, 15 (60%) were classified as medium difficulty, 7 (28%) as difficult, and 3 (12%) as easy. Items in the medium category had proportions between 0.40 and 0.70, indicating that approximately half of the participants answered correctly, an optimal range for diagnostic purposes (Negoro & Karina, 2019). The difficult items, including Questions 1, 2, 6, 13, 14, 15, and 18, had proportions between 0.20 and 0.30, suggesting that these concepts, including loanword spelling and word division, are particularly challenging and are strong candidates for revealing deep misconceptions. Items 7, 10, and 25 were classified as easy, reflecting students' relatively stronger grasp of those linguistic sub-topics.

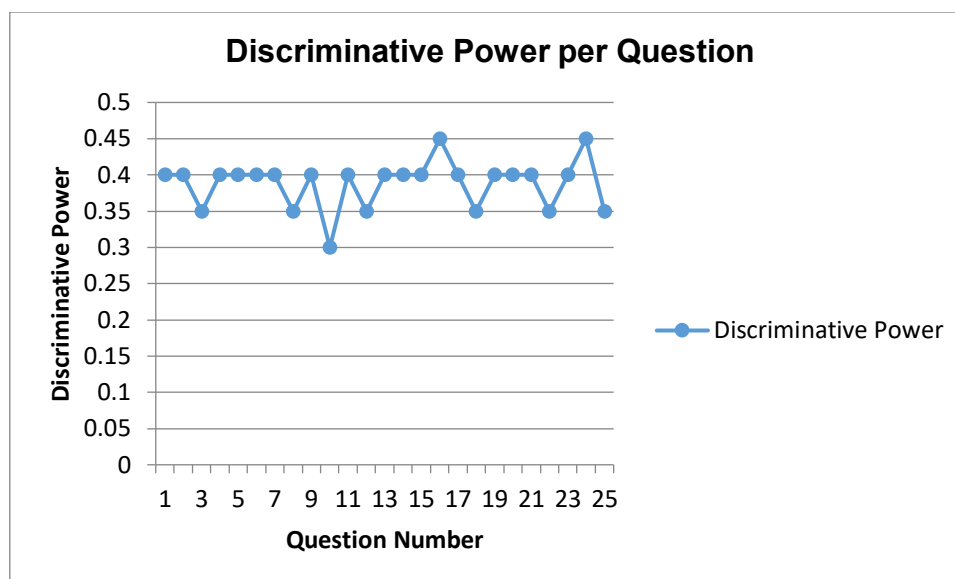


Figure 11. Analysis of discriminating power in the developed instrument

Based on the data obtained, the discrimination values of the items ranged from 0.33 to 0.73. Items with the highest values, such as item number 19, which reached 0.73, were classified as very good because they had the optimal ability to identify differences in ability among participants. In addition, most questions showed a discrimination index between 0.47 and 0.60. Questions such as numbers 1, 2, 3, and 5 fall into the "good" category because they are able to effectively classify participants based on their level of understanding of the material.

In contrast, there are also several items with lower discrimination values, although they are still within acceptable limits. Questions with values around 0.40, such as questions 4, 6, and 7, still show adequate performance even though their discrimination levels are not as high as other questions. In addition, questions with a value of 0.33, such as questions 12, 15, and 25, are categorized as "fair," which means that these questions can still be used but would be better if refined to be more effective in distinguishing participants' abilities (Rusilowati et al., 2024). Overall, this analysis shows that the majority of questions in the developed instrument have good discriminating power, although there are some questions that are recommended for revision to improve the accuracy and effectiveness of the evaluation.

Analysis of Students' Misconceptions in EYD V Material

To gain a deeper understanding of students' mastery of EYD V material, an analysis was conducted using the Four Tier Gamified Diagnostic Test instrument. The table below shows

the percentage of students who had misconceptions on each question, as indicated by incorrect answers, high confidence levels, and reasons that did not match the actual concept.

Question No.	Understands Concept		Does Not Understand the Concept		Misconceptions	
	Total	%	Number	%	Number	%
1	8	20	25	62.50%	7	17.50%
2	7	17.50%	22	55	11	27.50%
3	6	15	20	50	14	35
4	5	12.50%	19	47.50%	16	40
5	9	22.50%	21	52.50%	10	25
6	10	25	20	50	10	25
7	4	10	18	45	18	45
8	12	30	16	40	12	30
9	14	35	15	37.50%	11	27.50
10	11	27.50%	17	42.50%	12	30
11	5	12.50%	20	50	15	37.50%
12	7	17.50%	24	60	9	22.50%
13	8	20	22	55	10	25
14	10	25	21	52.50%	9	22.50%
15	12	30	18	45	10	25
16	11	27.50%	19	47.50%	10	25
17	9	22.50%	20	50	11	27.50%
18	6	15	22	55	12	30
19	7	17.50%	21	52.50%	12	30
20	5	12.50%	19	47.50%	16	40
21	8	20	22	55	10	25
22	7	17.50%	23	57.50%	10	25
23	6	15	20	50	14	35
24	9	22.50%	21	52.50%	10	25
25	10	25	20	50	10	25

Table 3. Misconception analysis data

Based on the data above, from 25 questions given to 40 students, it was found that the level of misconceptions varied between 17.5% and 45%. This fact illustrates that many students do not yet have a complete and correct understanding of EYD V rules. In fact, some students showed a high level of confidence in incorrect answers, reflecting conceptual misconceptions rather than mere ignorance.

The highest level of misconception was recorded in Question 7, with 18 out of 40 students (45%) experiencing misconceptions. In addition, questions 4 and 20 also showed high numbers, with 40% of students experiencing misconceptions in each question. This indicates that the concepts tested in these questions were among the most difficult to

understand, and were even misunderstood with certainty (Kaltakci-Gurel et al., 2017). These questions were most likely related to the use of punctuation, the spelling of foreign loanwords, or the rules for capitalization, which are often confusing and lead to misinterpretation.

Other questions, such as numbers 3, 17, and 23, recorded misconception rates of 35% each, reinforcing the indication that many students made mistakes in understanding the rules that should be basic in writing good and correct Indonesian. In addition, questions 11 and 18 recorded misconception rates of 37.5% and 30%, meaning that nearly one-third of respondents were confident in their incorrect answers.

However, there were also several questions with lower misconception rates, such as Question 1 (17.5%), Questions 8 and 24 (30%), and Questions 15 and 25 (25% each). This shows that students are relatively more capable of understanding and applying certain concepts correctly. This low level of misconception may be influenced by the frequent use of these concepts in everyday writing practice or by more effective learning methods.

CONCLUSION

This study successfully developed a Four-Tier Gamified Diagnostic Test (FT-GDT) instrument for EYD V material, comprising 25 multiple-choice items, each structured across four response tiers (answer, confidence in answer, reasoning, and confidence in reasoning), drawn from 15 core linguistic aspects of EYD V, including capitalization, punctuation, loanwords, compound words, and preposition spelling. Expert validation confirmed high feasibility, with average scores of 86.7% from subject matter experts and 88.3% from media experts. Empirical validity testing identified 16 valid items from the 25 tested ($r \geq 0.254$), while reliability analysis of 30 respondents yielded a Cronbach's Alpha of 0.79, indicating adequate internal consistency. Item difficulty analysis revealed that 60% of items fell in the medium range, confirming a balanced and diagnostically representative distribution. Misconception analysis involving 40 students demonstrated that misconception rates ranged from 17.5% to 45%, with many students expressing high confidence in incorrect answers, confirming that the observed errors reflect genuine conceptual misconceptions rather than simple knowledge deficiencies.

The principal contribution of this study lies in extending the four-tier diagnostic assessment paradigm, previously established primarily in the natural sciences, to the domain of Indonesian language education. By developing and validating an FT-GDT

instrument specifically for EYD V material, this study demonstrates that the four-tier diagnostic approach can effectively reveal the depth and nature of linguistic misconceptions that conventional assessments fail to detect. The integration of gamification through the Wordwall platform further distinguishes this instrument by enhancing student motivation and engagement during the assessment process, without compromising its diagnostic rigor. These findings provide Indonesian language educators with a practical, evidence-based tool for formative and diagnostic assessment, with direct implications for instructional design, curriculum development, and the targeted remediation of conceptual errors in Indonesian spelling instruction at the higher education level.

Nevertheless, this study is subject to several limitations that serve as directions for future research. The sample sizes were relatively small, 30 participants for reliability testing and 40 for misconception analysis, which limits the generalizability of the findings to broader student populations. Additionally, nine of the 25 items were found to be invalid and require systematic revision, particularly in their wording, distractor design, and alignment with EYD V sub-topics, before the instrument can be deployed at full diagnostic capacity. The Wordwall platform, while effective for gamification, imposed a technical constraint in the form of the absence of a dedicated pre-test instruction space, necessitating supplementary instruction delivery outside the platform. Future research should address these limitations by revising invalid items based on qualitative feedback, conducting large-scale trials with more diverse and geographically varied populations, and exploring the instrument's integration with adaptive digital learning platforms, such as Learning Management Systems or AI-assisted assessment tools, to expand its diagnostic depth and applicability in varied Indonesian language learning contexts.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support provided by the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia (Kementerian Pendidikan Tinggi, Sains, dan Teknologi/Kemendiktisaintek) through the 2025 research funding program. The authors sincerely appreciate this support, which made the completion of this research possible. We also express our gratitude to all individuals and institutions who contributed to the implementation of this study through their academic support, valuable insights, and assistance during the research process. Finally, we thank the

anonymous reviewers and editors for their constructive comments and valuable suggestions, which have significantly improved the quality of this manuscript.

DECLARATIONS

Ethical considerations

The authors declare that this study was conducted in accordance with accepted ethical principles for educational research. Participation in this study was voluntary, and informed consent was obtained from all participants prior to their involvement. Participants were informed of the purpose of the study and assured that their responses would be kept confidential and used solely for research purposes. The identities of all participants have been anonymized to protect their privacy.

Use of artificial intelligence (AI)

“The authors declare that the generative artificial intelligence (AI) tool [name of the tool] was used exclusively for language editing and/or grammatical improvement. The use of AI did not influence the scientific content, study design, data analysis, data interpretation, results, or conclusions of the manuscript. Full responsibility for the content remains with the authors.”

Conflict of Interest

The authors declare no conflicts of interest.

Funding

This research was financially supported by the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia (Kementerian Pendidikan Tinggi, Sains, dan Teknologi/Kemendiktisaintek) through the 2025 research funding program.

REFERENCE

- A. F, N. (2018). Cognitive load theory in the context of second language academic writing. *Higher Education Pedagogies*, 3(1), 385–402.
<https://doi.org/10.1080/23752696.2018.1513812>
- Abdurrasyid, M. (2024). Analisis Kesalahan Penggunaan Ejaan Bahasa Indonesia (EBI) pada Surat Kabar Pergeseran Media Arus Utama dan Meluasnya Ruang Media Baru. *Jurnal Pendidikan, Bahasa Dan Budaya*, 3(4), 203–208.
<https://doi.org/10.55606/jpbb.v3i4.4672>
- Aksoy, A. C. A., & Erten, S. (2022). A four-tier diagnostic test to determine pre-service science teachers' misconception about global warming. *Journal of Baltic Science Education*, 21(5), 747–761.
- Anderson, L. W., & Krathwohl, D. R. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman.
- Aristiawan, A., Wilujeng, H., Erwanto, M. R. Y., & Murtafiah, W. (2025). Development of a four-tier diagnostic test to detect middle school students' misconceptions in algebraic thinking. *Prima: Jurnal Pendidikan Matematika*, 9(2), Article 2.
<https://doi.org/10.31000/prima.v9i2.12700>
- Astuti, I. A. D., Bhakti, Y. B., Prasetya, R., & Rahmawati, Y. (2023). Four Tier-Relativity Diagnostic Test (4T-RDT) to Identify Student Misconception. *JIPF (Jurnal Ilmu Pendidikan Fisika)*, 8(1), 75. <https://doi.org/10.26737/jipf.v8i1.3668>
- Caleon, I. S., & Subramaniam, R. (2010). Do Students Know What They Know and What They Don't Know? Using a Four-Tier Diagnostic Test to Assess the Nature of Students' Alternative Conceptions. *Research in Science Education*, 40(3), 313–337.
<https://doi.org/10.1007/s11165-009-9122-4>
- Cirocki, A., Anam, S., Drajiati, N. A., & Soden, B. (2025). Assessment Literacy Among Indonesian Pre-service English Language Teachers: A Mixed-Methods Study. *Iranian Journal of Language Teaching Research*, 13(1), 23–44.
<https://doi.org/10.30466/ijltr.2025.55691.2851>
- Cressa, J., & Adni, D. F. (2023). Level Kognitif Taksonomi Bloom pada Soal Mata Pelajaran Bahasa Indonesia. *J-LELC: Journal of Language Education, Linguistics, and Culture*, 3(1), Article 1.
- Diani, R., Alfin, J., Anggraeni, Y. M., Mustari, M., & Fujiani, D. (2019). Four-Tier Diagnostic Test With Certainty of Response Index on The Concepts of Fluid. *Journal of Physics: Conference Series*, 1155, 012078. <https://doi.org/10.1088/1742-6596/1155/1/012078>
- Farida, H., Haryono, H., & Lestari, W. (2020). Development of Speaking Skill Assessment Instruments on Learning Indonesian Language Android-Based at Primary School. *Educational Management*, 9(1), Article 1.
- Fariyani, Q., Rusilowati, A., & Sugianto -. (2015). PENGEMBANGAN FOUR-TIER DIAGNOSTIC TEST UNTUK MENGUNGKAP MISKONSEPSI FISIKA SISWA SMA KELAS X. *Journal of Innovative Science Education*, 4(2), Article 2.
<https://journal.unnes.ac.id/sju/jise/article/view/9903>
- Febrina, R., & Yusuf, M. (2021). *The Effectiveness of Improving the Understanding of Indonesian Students Department of Indonesian Literature of Andalas University*

- Through the Language Card Playing Method*. 591–598.
<https://doi.org/10.2991/assehr.k.210202.102>
- Fратиwi, N. J., Kaniawati, I., Suhendi, E., Suyana, I., & Samsudin, A. (2017). The transformation of two-tier test into four-tier test on Newton's laws concepts. *AIP Conference Proceedings*, 1848(1). <https://doi.org/10.1063/1.4983967>
- Istiyono, E., Dwandaru, W. S. B., Fenditasari, K., Ayub, M. R. S. S. N., & Saepuzaman, D. (2023). The Development of a Four-Tier Diagnostic Test Based on Modern Test Theory in Physics Education. *European Journal of Educational Research*, volume–12–2023(volume–12–issue–1–january–2023), 371–385. <https://doi.org/10.12973/eu-ier.12.1.371>
- Jannah, M., Erniwati, E., Mongkito, V. H. R., & Yuris, M. (2022). Analisis Miskonsepsi Siswa Menggunakan Four Tier Diagnostic Test pada Pokok Bahasan Fluida Statis Siswa SMKN 2 Kendari. *Jurnal Penelitian Pendidikan Fisika*, 7(1), Article 1. <https://doi.org/10.36709/jipfi.v7i1.23084>
- Javidanmehr, Z., & Sarab, M. R. A. (2017). *Cognitive Diagnostic Assessment: Issues and Considerations*. 7(2).
- Joshi, R. M., Wijekumar, K., & Gillespie Rouse, A. (2022). International Perspectives on Spelling and Writing in Different Orthographies: Introduction to the Special Series. *Journal of Learning Disabilities*, 55(2), 83–86. <https://doi.org/10.1177/00222194211059836>
- Jumadi, J., Sukarelawan, M. I., & Kuswanto, H. (2023). An investigation of item bias in the four-tier diagnostic test using Rasch model. *International Journal of Evaluation and Research in Education (IJERE)*, 12(2), Article 2. <https://doi.org/10.11591/ijere.v12i2.22845>
- Kaltakci-Gurel, D., Eryilmaz, A., & McDermott, L. C. (2017). Development and application of a four-tier test to assess pre-service physics teachers' misconceptions about geometrical optics. *Research in Science & Technological Education*, 35(2), 238–260. <https://doi.org/10.1080/02635143.2017.1310094>
- Kessler, B. (2003). Is English Spelling Chaotic? Misconceptions Concerning Its Irregularity. *Reading Psychology*, 24(3–4), 267–289. <https://doi.org/10.1080/02702710390227228>
- Khairani, I., Gea, O. F., Panjaitan, K. A., Dewita, C. F., Manik, I. W., & Saragih, F. (2025). Analisis Kesalahan Kaidah Kebahasaan dalam Media Publikasi di Kota Medan: Penyebab dan Solusinya. *Pragmatik : Jurnal Rumpun Ilmu Bahasa Dan Pendidikan*, 3(2), 95–112. <https://doi.org/10.61132/pragmatik.v3i2.1548>
- Kirbulut, Z. D., & Geban, O. (2014). Using Three-Tier Diagnostic Test to Assess Students' Misconceptions of States of Matter. *Eurasia Journal of Mathematics, Science and Technology Education*, 10(5), 509–521. <https://doi.org/10.12973/eurasia.2014.1128a>
- Laila, L. M., & Nashoih, A. K. (2024). Taṭwīr Adawāt Taqwīm Ta'lim al-Balāghah 'abra Minaṣṣah Wordwall fī Ḍauī Madkhal al-Ikhtibār at-Tasykhīsyī li Arba'ati Mustawayāt: تطوير أدوات تقويم تعليم البلاغة عبر منصة وورد وول في ضوء مدخل الاختبار التشخيصي

- لأربعة مستويات. *LISANIA: Journal of Arabic Education and Literature*, 8(2), Article 2.
<https://doi.org/10.18326/lisania.v8i2.227-256>
- Latifah, U., & Damayanti, M. I. (2022). Pengembangan alat evaluasi pembelajaran menggunakan platform wordwall.net untuk siswa kelas ii sekolah dasar. *Jurnal Penelitian Pendidikan Guru Sekolah Dasar*, 10(06).
<https://ejournal.unesa.ac.id/index.php/jurnal-penelitian-pgsd/article/view/47490>
- Lin, N., Wu, H., Zheng, W., Liao, X., Jiang, S., Yang, A., & Xiao, L. (2024). IndoCL: Benchmarking Indonesian Language Development Assessment. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds), *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 4873–4885). Association for Computational Linguistics.
<https://doi.org/10.18653/v1/2024.findings-emnlp.280>
- Listyowati, I. (2025). Analisis Butir Soal Multiple Choice Mata Pelajaran Bahasa Indonesia Siswa Kelas 4 di SDN Bareng 5 Jombang. *Jurnal Review Pendidikan Dan Pengajaran*, 8(2), 3968–3976.
- Maison, M., Kurniawan, D. A., & Widowati, R. S. (2021). The Quality of Four-Tier Diagnostic Test Misconception Instrument for Parabolic Motion. *Jurnal Pendidikan Dan Pengajaran*, 54(2), 359–369. <https://doi.org/10.23887/jpp.v54i2.35261>
- Nabila, S., Repelita, T., Putra, P. A., & Risnawati, D. (2024). Analisis dampak globalisasi terhadap penggunaan bahasa indonesia dan peran pedoman umum ejaan bahasa indonesia (puebi). *Jurnal Pemasaran Bisnis*, 6(3), Article 3.
<https://journalpedia.com/1/index.php/jpb/article/view/2811>
- Nazriani, N., & Arsad, A. (2020). Kesalahan Penggunaan Bahasa Indonesia dalam Ruang Publik (Media Luar Ruang) di Kota Baubau. *Sang Pencerah: Jurnal Ilmiah Universitas Muhammadiyah Buton*, 6(1), Article 1. <https://doi.org/10.35326/pencerah.v6i1.552>
- Negoro, R. A., & Karina, V. (2019). Development of A Four-Tier Diagnostic Test For Misconception of Oscillation and Waves. *Jurnal Penelitian & Pengembangan Pendidikan Fisika*, 5(2), Article 2. <https://doi.org/10.21009/1.05201>
- Pahlefi, M. R. (2022). Pengembangan Instrumen Penilaian Keterampilan Menyimak (Mahārah al-Istima') dalam Pembelajaran Bahasa Arab. *Uktub: Journal of Arabic Studies*, 2(2), Article 2. <https://doi.org/10.32678/uktub.v2i2.6458>
- Park, C. (2012). Orthographic Differences and English L2 Spelling Errors. *Korean Journal of English Language and Linguistics*, 12(2), 285–302.
<https://doi.org/10.15738/kjell.12.2.201206.285>
- Prabowo, A., Herman, T., & Fatimah, S. (2023). *The Characteristics of Four-Tier Diagnostic Test: Classical Test Theory Perspective*. 7(1).
- Prabowo, A., Herman, T., Fatimah, S., & Ahniah, A. D. (2023). Development of Four-tier Multiple-choices Test to Diagnose Student's Misconception. *Kreano, Jurnal Matematika Kreatif-Inovatif*, 14(2), Article 2. (Junior HighSchool Student).
<https://doi.org/10.15294/kreano.v14i2.45400>
- Putri, N. P., & Handayani, N. (2024). *The Spelling Comprehension of PBSI Students STKIP PGRI Pacitan (Viewed from the Change of PUEBI to EYD V) | Proceeding International Conference on Learning and Education*.
<https://prosiding.stkippacitan.ac.id/index.php/icle/article/view/17>

- Rahman, G. F., Sofyan, D., & Gusnadi, G. (2017). Students Error In Using Capital Letters And Punctuation In Writing A Descriptive Text. *Pedagogia: Jurnal Ilmiah Pendidikan*, 9(1), Article 1. <https://doi.org/10.55215/pedagogia.v9i1.6667>
- Rahmawati, L. E., Wahyudi, A. B., Purnanto, A. W., Latifa, R., & Purnomo, E. (2022). Evaluasi Pelaksanaan Mata Kuliah Wajib Bahasa Indonesia di Perguruan Tinggi Muhammadiyah dan 'Aisyiah Menggunakan Model CIPP. *Imajeri: Jurnal Pendidikan Bahasa dan Sastra Indonesia*, 4(2), 92–102. <https://doi.org/10.22236/imajeri.v4i2.8763>
- Ramly, R., & Abduh, A. (2018). Exploring Cognitive Concepts in the National Assessment of the Indonesian Language. *The New Educational Review*, 53(3), Article 3.
- Ratnasari, C. I., & Mubasiroh, S. L. (2022). Evaluasi Hasil Pembelajaran Ejaan Bahasa Indonesia Sesuai EYD dan KBBI dengan Adopsi Teknologi. *Edumatic: Jurnal Pendidikan Informatika*, 6(2), Article 2. <https://doi.org/10.29408/edumatic.v6i2.6714>
- Rawh, P., Samsudin, A., & Nugraha, M. G. (2020). Pengembangan Four-Tier Diagnostic Test untuk Mengidentifikasi Profil Konsepsi Siswa pada Materi Alat-alat Optik. *WaPFI (Wahana Pendidikan Fisika)*, 5(1), 84–89. <https://doi.org/10.17509/wapfi.v5i1.22888>
- Rusilowati, A., Darsono, T., Sekarningtias, F. O., & Fariyani, Q. (2024). Development of A Four-Tier E-Diagnostic Test on The Topic of Momentum to Measure and Reduce Student Misconception. *Jurnal Pendidikan Fisika Indonesia*, 20(2), Article 2. <https://doi.org/10.15294/jpfi.v20i2.1297>
- Salele, N., Shahadat Hossain Khan, M., Hasan, M., & Ali, S. (2025). Advancing Four-Tier Diagnostic Assessments: A Novel Approach to Mapping Engineering Students' Conceptual Understanding in Microwave Engineering Course. *IEEE Access*, 13, 59886–59910. <https://doi.org/10.1109/ACCESS.2025.3555432>
- Sari, N. W. N., & Sunyono, S. (2019). Development Of The Three Tier Diagnostic Test Based 'Higher Order Thinking Skills' Instrument. *Dinamika Jurnal Ilmiah Pendidikan Dasar*, 11(2), 86–93. <https://doi.org/10.30595/dinamika.v11i2.5053>
- Setiadin, K., Repelita, T., Nurhaliza, S., Setiawan, N. D., & Auliya, Z. (2024). Penggunaan bahasa indonesia yang baik dan benar sesuai dengan puebi pada mahasiswa universitas buana perjuangan angkatan 2023. *Jurnal Pemasaran Bisnis*, 6(3), Article 3. <https://journalpedia.com/1/index.php/jpb/article/view/2827>
- Shahmirzadi, N. (2023). Validation of a Language Center Placement Test: Differential Item Functioning. *International Journal of Language Testing*, 13(1), 1–17. <https://doi.org/10.22034/ijlt.2022.336779.1151>
- Sugilastini, N. P. (2022). *Peningkatan pemahaman kaidah kebahasaan dalam puebi menggunakan media si gebi (sumber informasi gerai bahasa di instagram) komik strip sebagai materi ajar pada siswa kelas ix smpn 11 denpasar* [Masters, Universitas Widya Dharma]. <http://repository.unwidha.com:880/3140/>
- Tomasouw, J., & Marantika, J. E. R. (2023). Implementasi Pelaksanaan Entry Behavior, Matrikulasi dan Kemampuan Reseptif Mahasiswa (Suatu Studi Kasus pada Mahasiswa Program Studi Pendidikan Bahasa Jerman Tahun Ajaran 2022/2023). *J-*

- EDu: Journal - Erfolgreicher Deutschunterricht*, 3(1), Article 1.
<https://doi.org/10.30598/I-EDu.3.1.8-15>
- Welter. R., B., & Meredith. D, G. (2003). *Educational Research: An Introduction* (4th edn). Longman Inc.
- Widiyanto, A., Sujarwanto, E., & Prihaningtiyas, S. (2018). Analisis pemahaman konsep peserta didik dengan instrumen four tier diagnostic test pada materi gelombang mekanik. *Prosiding Seminar Nasional Multidisiplin*, 1, 138–146.
- Wirahyuni, K. (2022). *Pengembangan Instrumen Evaluasi Afektif Berbasis Digital Pada Pembelajaran Mata Kuliah Pengembangan Kepribadian Bahasa Indonesia Di Universitas Pendidikan Ganesha* [Doctoral, Universitas Pendidikan Ganesha].
<https://repo.undiksha.ac.id/9651/>
- Yang, D.-C., & Lin, Y.-C. (2015). Assessing 10- to 11-year-old children's performance and misconceptions in number sense using a four-tier diagnostic test. *Educational Research*, 57(4), 368–388. <https://doi.org/10.1080/00131881.2015.1085235>
- Zega, E. V. A. N., Ndruru, M., Harefa, N. A. J., & Bawamenewi, A. (2024). Analisis kesalahan bahasa indonesia di artikel jurnal taehao pbsi. *Jurnal Kata : Bahasa, Sastra, Dan Pembelajarannya*, 12(2 Sep), Article 2 Sep.